Image and Vision Computing 114 (2021) 104282

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Review article

A review of deep learning techniques for 2D and 3D human pose estimation

Miniar Ben Gamra *, Moulay A. Akhloufi

Perception, Robotics, and Intelligent Machines Research Group (PRIME), Dept of Computer Science, Université de Moncton, Moncton, NB, E1A 3E9, Canada

ARTICLE INFO

Article history:
Received 28 July 2021
Accepted 12 August 2021
Available online 18 August 2021

Keywords:
2D and 3D human pose estimation
Single-person and multi-person pose estimation
Deep learning
CNN
Computer vision

ABSTRACT

Inferring human pose from a monocular RGB image remains an interesting field of research in computer vision. It serves as a fundamental key for many real-world applications, including human-computer interaction, animation, and detecting abnormal or illegal human behavior. Despite the considerable progress made in this area during the last decade, the proposed methods face serious problems due to the huge variations in human appearance, occlusions, noisy backgrounds, viewpoints, and other factors that can change the context of the captured information. In this paper, we introduce a survey of state-of-the-art methods to highlight various research that have been proposed to tackle the 2D and 3D pose estimation tasks. Based on the number of persons in the image, two main pipelines are identified: single-person and multi-person methods. Each of these categories is divided into two groups according to the proposed architectures. Also, we provide a brief description of current datasets and the different metrics applied to evaluate the methods performances. Finally, we include a discussion about the advantages and disadvantages of the mentioned strategies.

Contents

* Corresponding author.
E-mail addresses: emb4506@umoncton.ca (M. Ben Gamra), moulay.akhloufi@umoncton.ca (M.A. Akhloufi).

https://doi.org/10.1016/j.imavis.2021.104282
0262-8856/

1. Introduction

Human pose estimation is one of the most important computer vision tasks in the past few decades. It tackles the task of automatically predicting and tracking human posture by localizing K body joints (also known as keypoints, such as elbows, wrists, etc.) in a given RGB image or video, as well as defining the orientation of its limbs. The potential of this task can be seen in a variety of applications in which it is involved whether to track human movement or to analyze and detect illegal or inappropriate human behavior. Pose estimation can also be applied in sports analysis to automatically track or assess human movement accuracy and serves as a fundamental tool in many other fields such as human-computer interaction and augmented reality.

In computer vision, there is a huge difference between the 2D and the 3D pose estimation. The 2D pose estimation consists of predicting the location of the body keypoints in a 2D space. In other words, the model estimates X and Y coordinates for each joint localization. Similarly, the 3D pose estimation infers the spatial position by adding an extra Z-axis to the predicted joint location. Typically, the 3D pose estimation is more challenging than 2D. Developing an accurate robust method requires a high computational complexity due to a variety of limitations, such as noisy background scenes, clothing, lighting conditions, small and barely visible joints, occlusions, and other factors that may change considerably the appearance of the body joints. In this review, we focus on both 2D and 3D human pose estimation fields.

Classical approaches address these difficulties using the pictorial structure framework that represents an object as a set of parts placed in a flexible configuration [1]. This strategy turns the pose estimation task into a tree-structured graphical model issue. Nevertheless, the above technique fails regarding non-visible limbs since it does not detect the correlation between body parts. Other works used the handcraft features including the Histogram of Oriented Gradients (HOG) features [2], contours, and edges. However, this technique does not generalize well, which decreases the model performance and makes it unsuitable for body joint localization.

Meanwhile, deep learning-based methods have achieved considerable advancement in several fields as they can capture the most relevant features within the metadata. Besides, the use of deep convolutional networks in recent human pose estimation methods, the current advancements in computing hardware, and most importantly, the availability of large-scale annotated datasets have contributed to achieving a remarkable improvement in terms of performance. Annually, novel techniques that outperform the old ones are proposed. The first work that shifted from classical approaches to deep learning was DeepPose proposed by Toshev et al. [3]. Currently, most of the proposed human pose estimators adopt ConvNets as their main backbone which has significantly improved the standard benchmarks.

It is also essential to keep in mind that there is a difference between inferring the pose of one person where its position is recognized, and localizing all joints of multiple persons where both their number and their positions in the image are unknown. We referred to these two pipelines respectively as a single-person and a multi-person pose estimation.

To estimate 2D single-person human pose, two directions have been followed. Some approaches consider this issue as a detection problem, which we called detection-based approaches, while others focus on regressing the body joint localization directly. The latter are called regression-based approaches. On the other hand, 2D multi-person approaches can be divided into two main categories based on their methodology: Top-down and bottom-up approaches. With the top-down approaches, and for a given image, the network applies in a first place a person detector to define a bounding box around each person instance, then locates the keypoints within each cropped area. The bottom-up models localize all present keypoints and then attempt to group them per person instance. Regardless of the adopted category, most deep learning models first introduce an encoder to extract features from the input image through a series of convolutional layers. We note that another category of methods uses depth information such as RGB-D to estimate human pose. However, in this paper, we focus only on methods that use RGB data.

When it comes to 3D methods, we define two distinct categories based on the chosen architecture. Each category has its advantages and its drawbacks. The first category, named one-stage approaches, regresses the 3D pose directly from the image features without adding any intermediate steps. In contrast, the second category, called two-stage approaches, derives from the fact that 2D and 3D poses could share common representations. It includes two steps where the first step consists of estimating 2D joint locations, typically using one of the top-performing 2D approaches, and the second step consists of recovering the 3D pose from the resulting 2D articulations. Usually, these approaches use both 2D and 3D datasets to enhance the model performances. Similarly, the 3D multi-person approaches present the same categorization as the 2D multi-person approaches. Fig. 1 illustrates the general taxonomy of this review.

With the introduction of deep learning-based methods, 2D single-person pose estimation methods have achieved very high performances. Recently, the focus has shifted to the multi-person pose estimation task, and more specifically in complex scenes where there are strong occlusions. In contrast, the 3D human pose estimation field

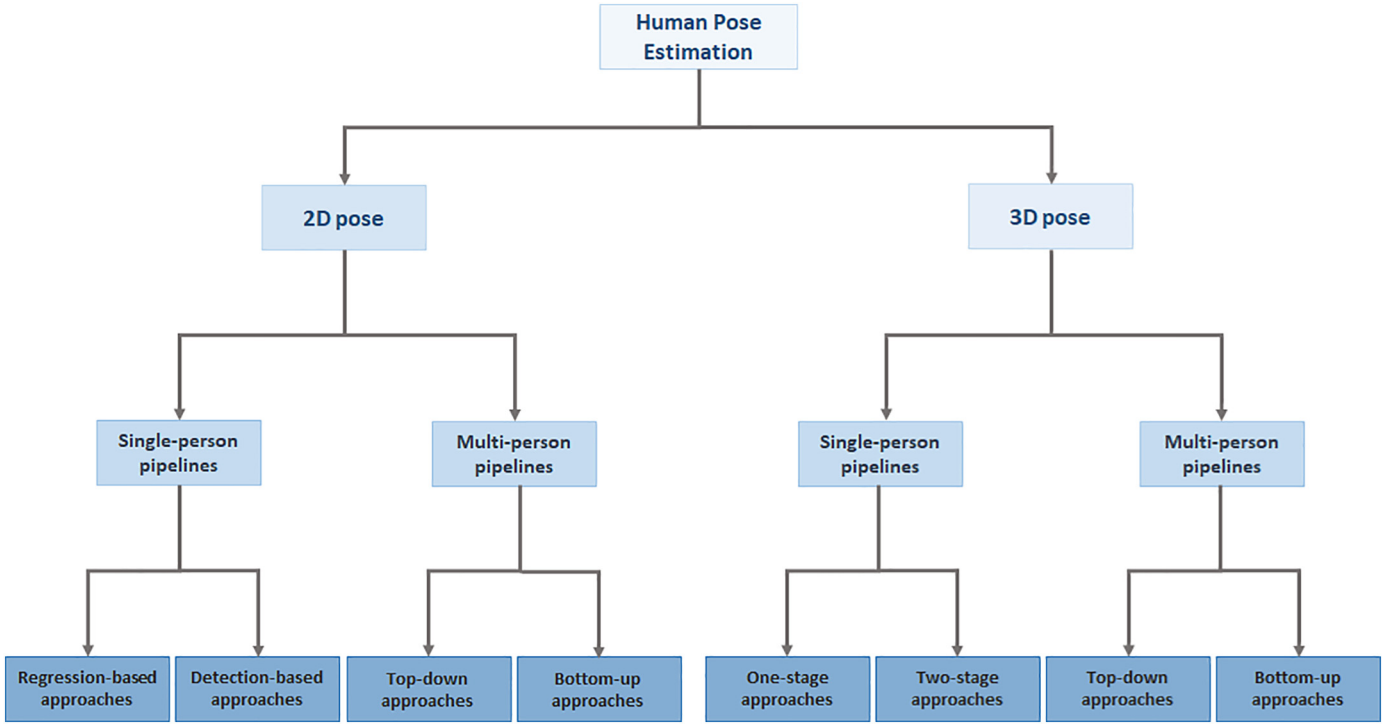

Fig. 1. The taxonomy of human pose estimation.

is still limited. The main reason behind this is the difficulty to recover the 3D ground-truth human joint position in an outdoor environment and, as a consequence, the lack of large-scale datasets that are annotated with the 3D human pose. Indeed, most 3D datasets are captured in a controlled lab environment, using motion capture systems (known as MoCap). Since variations in background, viewpoints and lighting are limited in such environment, the proposed methods do not generalize well in unconstrained environments.

Until 2016, most previous reviews focused on summarizing classical human pose estimation methods and did not provide any discussion about deep learning-based methods. Nowadays, we can say that the literature is quite developed since many studies have covered this area. Among them, we mention the survey realized by Sarafianos [4] that reviewed 3D human pose estimation models from RGB inputs and classified them as generative, discriminative, or hybrid methods. Li [5] provided a brief review of the 2D multi-person pose estimation methods. Dang et al. [6] provided a comparison among several 2D deep learning-based methods. Munea et al. [7] included a 2D review, as well as an interpretation of the proposed models. Recently, Zheng et al. [8] presented both 2D and 3D methods in a review. Despite their important contributions, these reviews do not include some recently published work.

The main objective of this paper is to address the shortcomings of previous surveys. Moreover, we provide a summary of recent 2D and 3D deep learning-based studies that address the main human pose estimation challenges, reveals the proposed solutions, and identifies the unsolved problems. In other words, this review may serve as a guideline for researchers interested in this field. Besides, the included discussion offers a valuable comparison of existing strategies and points to some possible future research directions. Furthermore, here are the key points of this review:

- Describe the different used 2D and 3D benchmarks, as well as the related evaluation metrics.
- Provide a summary of up-to-date deep learning-based models that address both 2D and 3D human pose estimation from RGB inputs. We classify the mentioned methods based on the proposed architectures.
- Compare the performances of the mentioned methods according to their categories and discuss the advantages and the limitations of the followed strategies.

2. Datasets

Having access to a large-scale dataset that offers a variety of examples is one of the most important steps in computer vision. Regarding the 3D human pose estimation task, the first encountered issue is the availability of large-scale outdoor datasets. Unlike 2D human pose datasets which offer a wide variety of benchmarks taken in both indoor and outdoor environments, most existing 3D human pose datasets are captured in a lab environment using MoCap systems. Unfortunately, this labeling process may limit variations in background, viewpoints, and lighting. In this section, we cover the most popular 2D and 3D human pose datasets.

2.1. MPII human pose

Max Planck Institute for Informatics (MPII) Human Pose dataset [9] is the standard benchmark for 2D human pose estimation. It contains around 25 K images with 40 K subjects. The dataset provides annotations for 16 body joints with various challenging camera directions. Images were collected from YouTube videos covering 410 daily human activities with complex poses, different scale variations, and image appearances. The dataset is split into three parts: 22 K for training, 3 K for validation, and 7 K for testing. The MPII test set provides various annotations such as body part occlusions and 3D torso and head orientations. MPII Human pose dataset has been used for single-person pose estimation models as well as for multi-person pose estimation models where it provides 3844 training samples and 1758 testing samples with both occluded and overlapped people. Fig. 2 shows some annotated examples from the MPII Human Pose dataset.

Fig. 2. Annotated examples from the MPII Human Pose dataset.

2.2. AI Challenger Human Keypoint Detection dataset

AI Challenger Human Keypoint Detection (AIC-HKD) [10] is a subset of a very large-scale dataset. Currently, it provides the largest training dataset offering 300 K images with 2D pose annotations. The dataset is split into three subsets where 210 K images are used for training, 30 K for validation, and 60 K for testing.

2.3. Leeds Sports Poses (LSP)

The LSP dataset [11] provides 2 K annotated images retrieved from Flickr after applying a search with different sports tags such as baseball, parkour, tennis, athletics, and so on. This kind of sports activity image is quite challenging in terms of appearance and joint localization. Moreover, images have been scaled to make the height of the majority of people equal to 150 pixels, which adds extra difficulty to the pose estimation task. Each person is labeled with the 2D coordinates of 14 body joints. Left and right joints are systematically annotated from a person-centered angle. The 'LSP Extended' (LSPET) [12] dataset adds around 10 K more training images to the LSP dataset.

2.4. COCO dataset

The COCO dataset [13] includes more than 200 K images with 250 K person instances. The provided annotations present the 2D poses of 17 body joints. The dataset is split into train/val/test-dev sets with respectively 57 K, 5 K, and 20 K images. These images, as shown in Fig. 3, illustrate a variety of human poses, unconstrained environments, different body scales, and occlusion patterns. Indeed, the majority of people in the COCO set are at medium and large scales.

2.5. Frames Labeled In Cinema (FLIC)

The FLIC dataset [14] consists of around 5 K images collected by running a state-of–the-art person detector on popular Hollywood movies as shown in Fig. 4. Each person instance is annotated with the 2D poses of 10 upper-body joints. 1016 images are allocated for testing. Most subjects are in front of the camera, which reduces occlusions between keypoints.

2.6. Look into Person (LIP)

The LIP dataset [15] is a large-scale single-person dataset that provides both human 2D pose localization and parsing annotations. It is composed of 50 K images divided into three subsets: 30 K for training, 10 K for validation, and 10 K for testing. Every labeled sample presents 16 body joints, as well as annotations for 19 semantic body parts with one background category. Images may include the whole body, half body, or just a part of the body. As illustrated in Fig. 5, images are gathered from real-world scenarios displaying humans in challenging poses and views, significant occlusions, various appearances, and low resolutions.

2.7. The extended PASCAL-Person-Part

The extended PASCAL-Person-Part dataset [16] includes the PASCAL VOC 2010 dataset [17] with extra annotations. It provides multi-person images with both 2D pose and analysis annotations for 14 joints and 6 body parts (i.e., head, torso, upper/lower arms, and upper/lower legs). In total, it contains 3533 images divided into two subsets: 1716 for training and 1817 for testing.

2.8. CrowdPose dataset

The CrowdPose dataset [18] is one of the most recent 2D human pose datasets that is captured in crowd and occlusion scenes. It is composed of around 20 K images selected from 30 K other images based on the crowd index (a metric that defines the level of the crowd in images). A total of 80 K human instances are annotated with the poses of 14 body keypoints. The training, validation, and testing sets contain respectively 10 K, 2 K, and 8 K images.

2.9. Penn Action

Penn Action Dataset [19] is a large-scale dataset that contains 2326 video sequences of 15 different actions, 1258 are used for training and

Fig. 3. Annotated examples from the COCO dataset.

1068 for testing. Each video clip is composed of 70 frames on average. Each frame is annotated with 2D poses of 13 body joints including head, shoulders, elbows, wrists, hips, knees, and ankles, in addition to another label that indicates whether a joint is visible or not. Fig. 6 shows examples from the Penn Action dataset.

2.10. Human3.6 M

Human3.6 M dataset [20] is currently the largest publicly available dataset for 3D human pose estimation. It contains about 3.6 million video frames for 11 subjects (5 females and 6 males) performing 17 daily activities such as eating, sitting, walking, making a phone call, and taking a photo. Videos are captured from 4 different views with RGB cameras, and only seven subjects are annotated using the MoCap system. The 2D poses can be obtained by projection with the known intrinsic and extrinsic camera parameters. Such a large dataset makes it possible to train data-driven pose estimation models. Fig. 7 shows examples from the Human3.6 M dataset.

2.11. HumanEva-I

HumanEva-I [21] is a smaller 3D pose dataset compared to Human3.6 M that has been widely used to benchmark previous work over the last decade. It contains three RGB video sequences recorded from three camera views at 60 Hz and synchronized with 3D body

Fig. 4. Examples from the FLIC dataset.

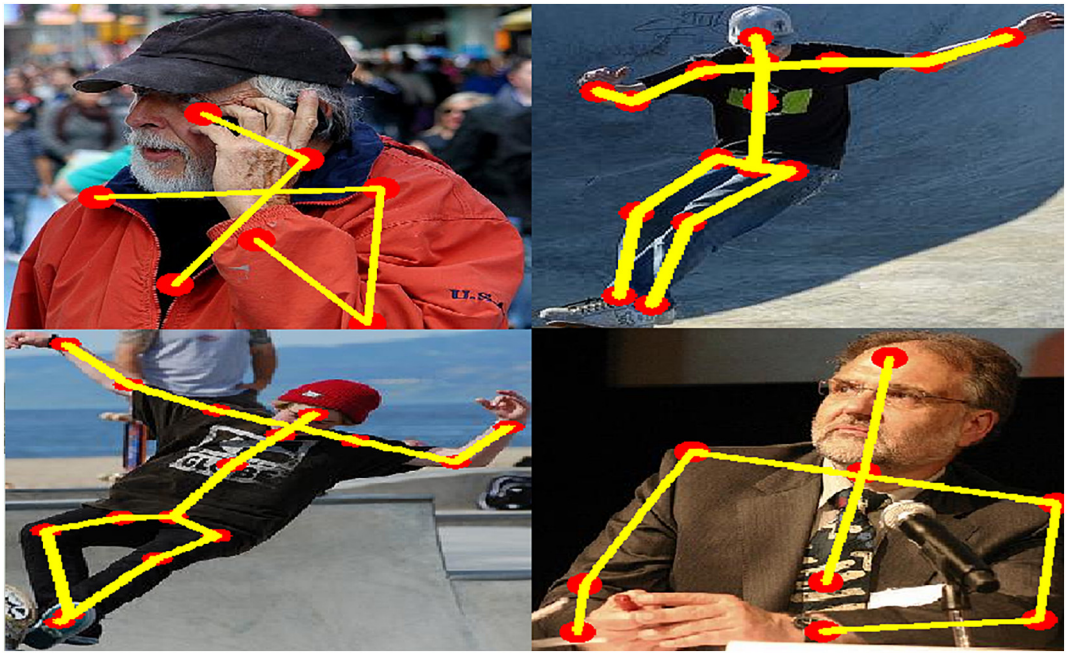

Fig. 5. Annotated examples from the LIP dataset.

poses obtained from a motion capture system. Fig. 8 shows examples from the HumanEva-I dataset.

2.12. MPI-INF-3DHP

MPI-INF-3DHP [22] is a recently proposed 3D dataset. Images were captured by a MoCap system with 12 synchronized cameras in both indoor and outdoor settings. It contains 8 subjects with diverse clothing.

2.13. Other datasets

Parse dataset [23] is a small 2D pose estimation dataset that provides 100 images for training and 205 images for testing. In addition to the 14 body joint annotations, it includes extra annotations such as facial expressions, gaze direction, and gender. The PoseTrack dataset [24] is a large-scale video-based dataset for 2D human pose estimation. It consists of 1356 video sequences, with 46 K annotated video frames, and

Fig. 6. Examples from the Penn Action dataset.

Fig. 7. Examples from the Human3.6 M dataset.

276 K body pose annotations. The BBC Pose dataset [25] is also a video-based dataset composed of 20 videos collected from the British Broadcasting Corporation (BBC) with the presence of a British Sign Language (BSL) signer. It contains 2D annotations for 7 human upper body keypoints and includes 610115 images for training, 309171 for validation, and 309260 for testing.

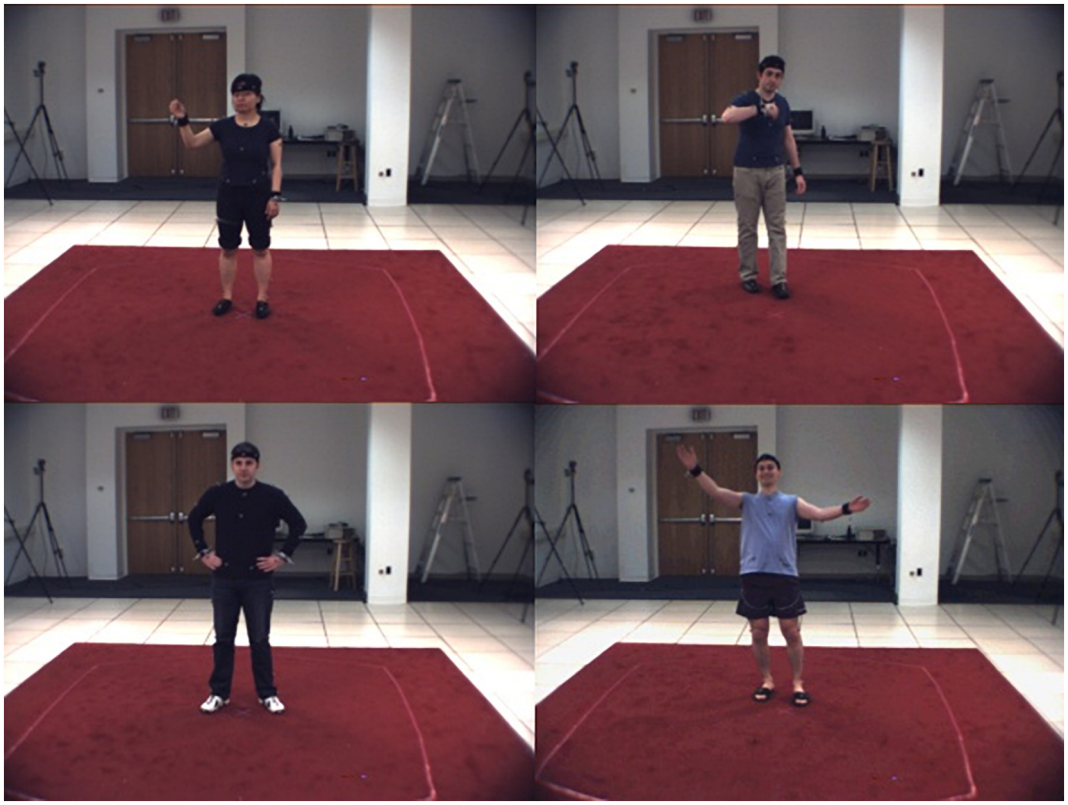

Fig. 8. Examples from the HumanEva-I dataset.

An extended BBC dataset [26] that added 72 additional training videos is also introduced. JHMDB (Joint-annotated Human Motion Data Base) [27] is another video-based dataset that provides 33183 frames annotated with 2D poses. HSSK (Human Skeletal System Keypoints) [10] contains 300 K images. For 700 K persons, 2D annotations of 14 keypoints are provided. MuPoTs-3D (Multi-person Pose estimation Test Set in 3D) [28] is composed of over 8 K images collected from 20 real-world scenarios involving up to three different subjects. It includes annotations of 14 body joints. CMU Panoptic [29] is a large-scale labeled dataset that provides multi-view 3D pose annotations with several social activities. From a total of 65 videos, only 17 videos present multi-person scenes and include the camera settings. The Campus Dataset [30] is a 3D pose dataset that presents three subjects interacting in an outdoor environment and captured by three different cameras. JTA (Joint Track Auto) [31] is a large dataset that tracks people in urban scenarios captured using a photorealistic video game. Five hundred and twelve full-HD videos are collected, where 256 are used for training and 256 for testing. The 3D Poses in the Wild dataset [32] is a video-based dataset that offers accurate 3D human poses. Both 2D and 3D annotations are provided for 60 video sequences captured from a moving phone camera.

3. Evaluation metrics

To evaluate a pose estimation model, the chosen metric has to consider several factors and features (e.g., upper/full human body joints, single/multiple pose estimation, human body scale). Consequently, numerous metrics have been developed to assess both 2D and 3D pose estimation methods. In this section, we mention the most well-known metrics.

3.1. Percentage of Correct Parts (PCP)

PCP [33] (more commonly referred to as PCP@0.5) is used to measure the rate of both 2D and 3D limb detection. A limb is considered as correctly located if the distance between two predicted joints and their ground truth localization is less than half of the limb length. The major drawback of this metric is that it penalizes shorter limbs more, as the thresholds are smaller for short body parts. For this reason, recent approaches have adopted other metrics.

3.2. Percentage Correct Keypoints (PCK)

The PCK metric [9] refers to the percentage of correct keypoint detection that falls within a specific distance. This distance can be considered as α pixels of the ground truth localization. For instance, PCK@0.5 refers to the PCK score with a threshold $\alpha = 50\%$. PCKh is similar to PCK, except that the distance is proportional to the head size. PCK is used for both 2D and 3D poses where it is known as 3DPCK.

3.3. Percentage of Detected Joints (PDJ)

The PDJ metric [3] follows almost the same rule as the PCK metric. Nevertheless, the detected joint is considered correctly located if it falls within a certain fraction of the torso length compared to the ground truth joint location. Thus, PDJ has been proposed as an alternative to the PCP metric to tackle the issue related to short limbs caused by using the same error threshold for all limbs. Usually, it is used to evaluate the 2D pose estimation models.

3.4. Object keypoint similarity

The OKS metric [13] shares the same objective of the Intersection over Union (IoU) metric in object detection tasks. It computes the distance between the predicted and the ground truth joint localization. The latter is normalized by the person scale, as illustrated in eq. 1, where d_i indicates the Euclidean distance between the predicted joint and its ground truth joint, v_i is the ground truth visibility indicator, s is the person scale, and k_i is a keypoint constant that reflects the fall-off.

$$\frac{\sum_i exp\left(-d_i^2/2s^2k_i^2\right)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{1}$$

3.5. Average Precision (AP) and Average Recall (AR)

The AP [13] is the primary COCO challenge metric that reflects the mean AP of 10 positions (with OKS = 0.50:0.05:0.95). AP^{50} and AP^{75} have also been mentioned in several works where the OKS is respectively equal to 50% and 75%. AP Across Scales is another form of this metric where AP^{Medium} is used to evaluate the model performance on medium objects (the object area is between $32^2$and 96^2 and AP^{Large} is used for large objects. Similarly, AR^{50}, AR^{75}, AR^{Medium}, and AR^{Large} [13] are also introduced. In addition, the mean Average Precision (mAP) over all classes is largely used to evaluate methods that involve the MPII and PoseTrack datasets.

3.6. Human3.6 M dataset evaluation metrics

To measure the 3D pose estimation method performances on the Human3.6 M dataset, several evaluation protocols have been followed. The first protocol is the most frequently used. It consists of using subjects 1, 5, 6, 7, 8 for training and subjects 9 and 11 for testing. The evaluation metric is called MPJPE as the Mean Per Joint Position Error and is calculated in millimeters. It indicates the mean Euclidean distance between the predicted and the ground truth joint localization. The second protocol indicates the prediction error after applying an alignment of translation, rotation, and scaling on the ground truth. It introduces 6 subjects (S1, S5, S6, S7, S8, and S9) for training and only subject S11 for testing. The error, called P-MPJPE, is averaged over 14 joints. The last protocol includes the same training and testing subjects as Protocol No. 1. However, the evaluation is limited to sequences captured by the front camera ("cam 3") from Trial 1 and the original video is not subsampled. The error measure, named N-MPJPE, used the 3D pose error as described in protocol No. 2. The protocol aligns the predicted poses with the ground truth only in scale. Unlike classic metrics, a better model has to achieve a lower score.

3.7. Area Under the Curve (AUC)

Area Under the Curve (AUC) metric measures the whole range PCK thresholds (e.g., when α runs from 0 to 0.5). It tracks the ability of the model to distinguish each body joint.

4. 2D human pose estimation

The 2D pose estimation aims to infer human joints localization in images or videos. Understanding a person posture through its pose serves as a fundamental tool in several research areas such as human-machine interaction, animation, and action recognition. As well as many other vision tasks, this field has been significantly improved through the introduction of Deep Convolutional Neural Networks (DCNNs). However, it still remains a very challenging task. A good pose estimation method has to be robust to human appearance variations such as clothing and occlusion, and invariant to any possible deformation that may change the context of the extracted features.

To address these challenges, early work used robust image features and sophisticated structured predictions. Based on the number of individuals in the input image, pose estimation approaches can be categorized as a single-person or a multi-person pipeline. Usually, it is easier

to detect the pose for one person. Otherwise, multi-person approaches aim to estimate poses for all individuals in a given image. For both categories, an interesting advancement is achieved, particularly due to the availability of the standard benchmarks and the high interest to introduce this technology in various applications. In this section, we discuss in detail some recent single-person and multi-person approaches.

4.1. Single-person pipelines

Images captured from real-world activities may include many people in the same scene. To this end, most single-person approaches involve a person detector to crop the region that contains one individual. On the other hand, inferring human pose can be considered either a regression problem or a detection problem. Regression-based methods tackle mapping directly the input image to body keypoint positions. However, detection-based approaches typically attempt to detect keypoints individually by producing keypoint heatmaps and then aggregate them in post-processing steps to generate the final predicted pose. In the following, we discuss in detail the difference among each category, as well as the recently proposed approaches.

4.1.1. Regression-based approaches

To predict joint localization, a lot of works follow the regression-based paradigm. DeepPose [3], introduced by Toshev et al., was the first work that proposed to use Deep Neural Networks to implicitly capture the full-body context. This study formulated the pose estimation as a regression problem towards body joints. Carreira et al. [34] proposed a framework that enhances the feature extractors to capture both input and output domains, by following a novel process, named iterative error Feedback (IEF). Usually, final results are predicted in one shot, unlike IEF, which presents a self-correcting model that progressively adjusts an initial solution by returning error prediction and rectify it iteratively.

Most previous regression-based methods were not as efficient as detection-based methods for estimating human pose. One of the contributing factors behind this is that joint dependencies are not well exploited. Therefore, Sun et al. [35] proposed a structure-aware approach, named Compositional Pose Regression. Since bones are more stable and easier to learn than joints, the proposed approach involves a novel reparameterized pose representation that uses bones instead of joints. Thus, the network engages a joint connection structure to define a compositional loss function that encodes the long-range interactions between bones.

The argmax operation used in detection-based pipelines has a lot of drawbacks. To this end, Luvison et al. [36] presented a new regression-based method that integrates the soft-argmax operation to convert feature maps directly to body joint coordinates. In the proposed architecture, contextual information is directly accessible and is easily aggregated to the final predictions. Moreover, this method does not require heatmap generation during the training phase.

Coordinate decoding refers to the operation that transforms the generated heatmap to joint coordinates, which constitutes the final prediction. The coordinate encoding is the reverse process that uses the original image as input. Unlike current research that focuses on designing more effective CNN structures, Zhang et al. [37] highlighted the important role of the joint coordinate representation by studying its various axes including the encoding and decoding process. They formulate a novel Distribution-Aware coordinate Representation of Keypoints (DARK) method with efficient Taylor-expansion-based coordinate decoding, and unbiased sub-pixel centered coordinate encoding.

4.1.2. Detection-based approaches

In detection-based approaches, the ground truth is generated from joint positions, usually by applying a 2D gaussian distribution centered on the joint location. After the introduction of convolutional neural networks, the contextual information is captured using convolutional layers. The majority of the suggested approaches are derived from the Stacked HourGlass network [38] which is described in detail in Section 4.2.

Previous state-of-the-art approaches address two main challenges. The first challenge is to create a keypoint heatmap by predicting the probability that each pixel may be a joint. The second challenge concerns spatial refinement, which consists of refining the resulting joint confidence by leveraging the spatial configuration of the human body. In the study realized by Sun et al. [39], a novel spatial configuration refinement algorithm that reduces human pose variations is introduced. The diversity of body part orientations is the main factor causing the huge variations in joint location. To this end, a human body and limb normalization scheme are proposed to reduce this kind of diversity and consequently generate compact distributions. Both multi-scale supervision and multi-scale fusion are also introduced to improve the joints localization process on different resolutions. Ke et al. [40] presented a robust multi-scale structure-aware approach that addressed the drawbacks of the recent version of HourGlass models by combining multi-scale feature combination, multi-scale supervision, information scheme, structure-aware loss and a keypoint masking training method to boost the pose estimator performances for occluded keypoints in complex or crowded environments.

Pyramid methods are widely used in DCNN thanks to their important impact in dealing with scale changes during the inference phase. However, they are not yet well explored in the pose estimation models. Inspired by the HourGlass network, Yang et al. [41] suggested a new framework, named Pyramid Residual Modules (PRMs). Based on the input multi-scale features, the PRMs learn convolutional filters. Besides, the utilization of residual units in the HourGlass network is problematic. More specifically, the summation of two residual unit outputs approximately doubles the output variance which affects the optimization process. To address this issue, they also proposed a simple but efficient scheme with negligible additional parameters.

Generative adversarial networks [42] are also used to estimate human pose. Chou et al. [43] employed a generative adversarial network where the generator and the discriminator share the same architecture. For both components, Stacked HourGlass is used as a network. The generator produces keypoint heatmaps based on the image features, and the discriminator aims to distinguish real heatmaps from fake ones and give the generator useful hints to improve the heatmap generation. After the training phase, the discriminator is excluded and only the generator is used as a pose estimator. Chen et al. [44] proposed an Adversarial PoseNet network. The network includes a multitask pose generator with two discriminator networks. The generator is designed to simultaneously regress the pose and the occlusion heatmaps. Both discriminators are designed to predict the probability of classifying the generated poses on real or fake poses. Using the pose and the occlusion heatmaps, the pose discriminator verifies whether the generated pose satisfies the joint connectivity constraints as well as if it is close to the ground truth heatmaps. The second discriminator is used to predict the generated pose heatmap confidence. Using this strategy, the network learns to be more robust to occlusions, overlapping, and twisting of human bodies. Wang et al. [45] focused on the problem caused by data corruption such as blur and pixelation and presented robust benchmarks named COCO-C, MPII-C, and OCHuman- C. They proposed AdvMix, a novel adversarial data augmentation algorithm implemented with knowledge distillation. The network helps improve the existing pose estimator performance under severe corruption.

Human parsing representation can also provide useful information for the human pose estimation task. Nie et al. [46] proposed a novel framework that uses the human parsing information to predict joints localization. The implemented network, named Parsing Induced Learner, refines inaccurate localization and corrects false body joints classification. PIL is composed of a parsing encoder and a pose model parameter adapter. The encoder transforms an input image into high-level parsing representations and the adapter learns to adjust the pose model

parameters by leveraging parsing representations. The end-to-end model can be used for both single-person and multi-person pose estimation.

The cascaded deep networks are other interesting architectures. Su et al. [47] proposed a novel Cascade Feature Aggregation (CFA) approach where the HourGlass blocks are replaced by ResNet. Features extracted from different stages are combined to obtain both local and global context information which makes CFA more robust to variations including illuminations and partial occlusions. To control the data flow for each channel, Bulat et al. [48] redesigned residual connections and proposed learnable channel-wise soft gated skip connections. Moreover, they introduced a hybrid network that fused the HourGlass and U-Net architectures. This network minimizes the number of identity connections to reduce the model size and complexity while maintaining a high accuracy. The work proposed by Artacho and Savakis [49] introduced UniPose, a one-stage unified framework that aims to localize human joints and person bounding boxes without any post-processing. The network is based on "Waterfall" Atrous Spatial Pooling (WASP) module that increases the receptive field of the network by combining the cascaded approach for Atrous Convolution with the larger FOV obtained from the parallel configuration of the Atrous Spatial Pyramid Pooling (ASPP) module. The proposed method presents a better interpretation of the contextual information and helps to produce more accurate pose estimation. Fig. 9 shows the Waterfall architecture in the WASP module. The study also proposed UniPose-LSTM that adopts a linear sequential LSTM configuration to estimate pose in video sequences.

4.2. Multi-person pipelines

Inferring the pose of all persons in a given image presents several challenges, especially if we are talking about a crowded scene. In fact, an image may contain an unknown number of people at various positions and scales. Moreover, the interaction between people causes several prediction issues that are mainly related to occlusions, which make the joint detection and the body part association quite difficult. For real-time applications, the computational time increases considerably with the number of people in the image, which has a direct impact on the performances of these systems. Consequently, all these factors must be taken into consideration to achieve a high accuracy score during inference.

Multi-person pose estimation can be addressed by following two types of strategies. The first strategy, called Top-down, consists of using a person-detector to define the region that contains a single person, and then apply a pose estimator at its output. The second strategy, called bottom-up, consists of first identifying all keypoints, then associating the resulting joints by person instance. In this section, we describe the most recent approaches that follow each paradigm.

4.2.1. Top-down methods

A top-down approach involves two main elements: a human detector and a single-person pose estimator. Most research uses existing human detectors such as Faster R-CNN [50], Mask R-CNN [51] and FPN [52]. DeepPose [3] is the first top-down approach that introduced deep neural networks to infer the human pose. First, a face-based person detector is used to estimate the person position, and then a multi-step cascaded network is used to directly regress the joint localization. Wei et al. [53] proposed Convolutional Pose Machines (CPM), the first end-to-end pose estimation model. CPM aims to generate progressively accurate joint location through a multi-stage convolutional network, without requiring an explicit graphical style inference. At each stage, the image features and the belief maps produced in the previous stage are used as input. The belief maps provide the subsequent stage an expressive non-parametric encoding of the spatial uncertainty of location for each part, allowing the CPM to learn rich image-dependent spatial models of the relationships between parts. To avoid gradient vanishing, intermediate supervision is used after each stage. Newell et al. [38] proposed Stacked HourGlass which is one of the most important networks that has been used in several works as a backbone. It involved repeated bottom-up and top-down processing applied with intermediate supervision to estimate human pose. As illustrated in Fig. 10, the HourGlass network uses consecutive pooling layers to get a very low resolution, then uses up-sampling layers and combines features across multiple resolutions by skip connections. The network has shown high robustness against various challenges, such as strong occlusions and the presence of multiple people in proximity. Papandreou et al. [54] proposed a two-stage pipeline to estimate 2D human pose in the 'wild'. First, they introduced Faster R-CNN [50] on top of a ResNet-101 [55] CNN as a person-detector. Second, the joint locations of each proposed person box are predicted using ResNet applied in a fully convolutional network to define activation heatmaps and offsets for each keypoint. To combine these outputs, a novel aggregation procedure is introduced to obtain highly localized keypoint predictions. Finally, a novel keypoint-based

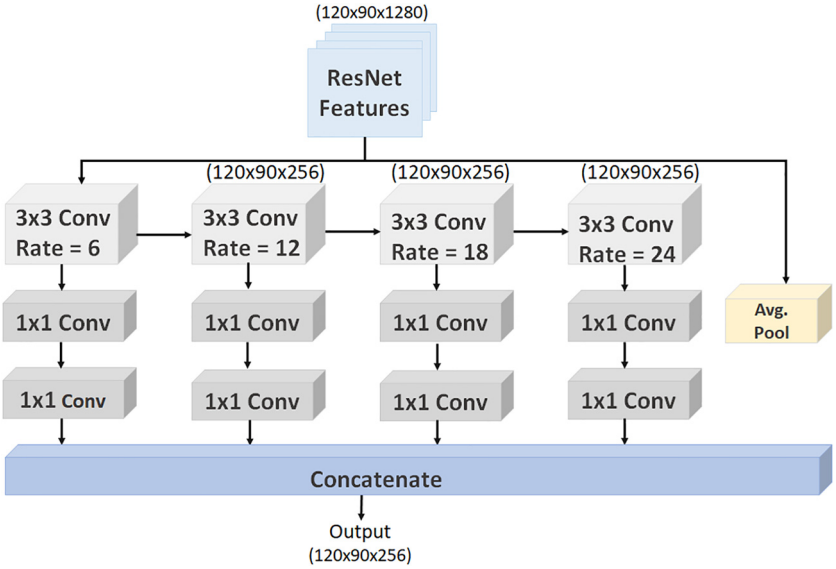

Fig. 9. Waterfall architecture in the WASP module.

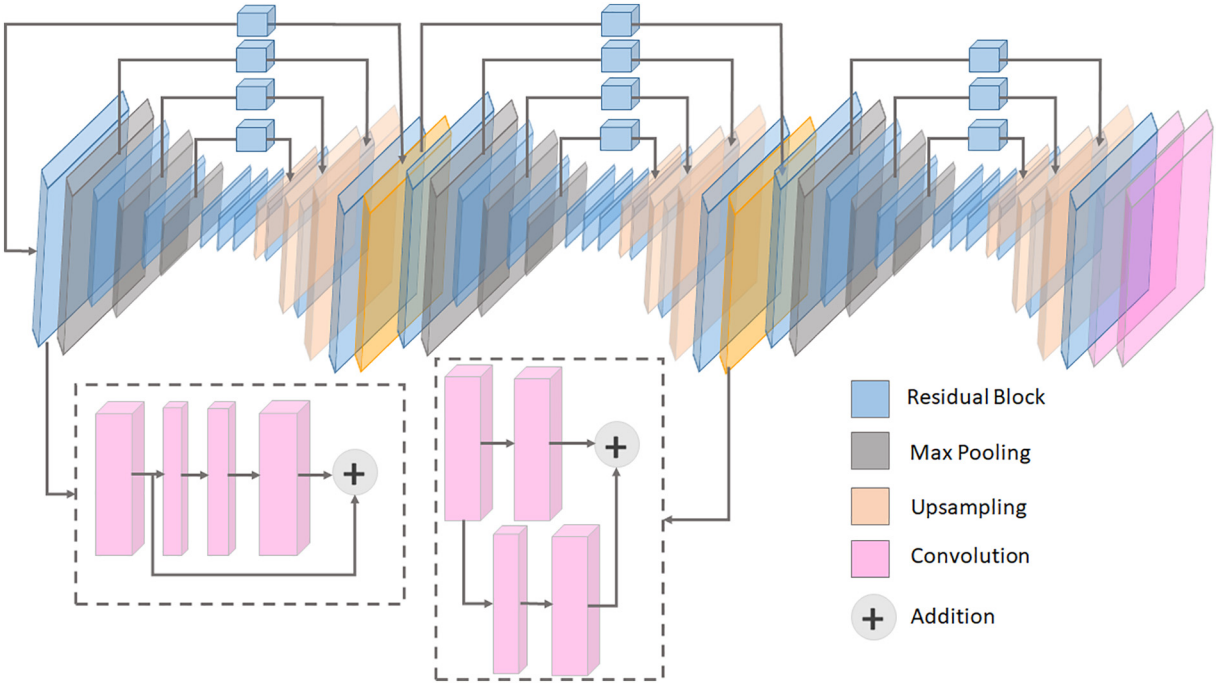

Fig. 10. The architecture of 3-Stack HourGlass.

Non-Maximum-Suppression (NMS) mechanism built directly on the OKS metric, called OKS-NMS, is proposed to eliminate duplicate pose detections. They also proposed a novel keypoint-based confidence score estimator.

Unfortunately, top-down methods are highly influenced by the person detector performances. Errors produced by the bounding box detection can directly affect the pose estimation process. To address this issue, Fang et al. [56] proposed a novel regional multi-person pose estimation (RMPE) framework to ensure an accurate pose prediction even for inaccurate predicted human bounding boxes or redundant detections. The framework is mainly composed of three components: the first is a Symmetric Spatial Transformer Network (SSTN) which is connected to the SPPE to extract a high-quality single-person region from an inaccurate bounding box. For the redundant detection issue, a parametric pose NMS scheme is included. The final component involves a new pose distance measure to compare the similarity of poses to eliminate redundant ones. In the study proposed by Chen et al. [57], a human detector is implemented based on FPN [52] to generate human bounding boxes, then a novel two-stage network, named Cascaded Pyramid Network (CPN), is introduced to detect "easy" and "hard" keypoints. The first stage, called GlobalNet, is a feature pyramid network that uses ResNet as a backbone to estimate "simple" keypoints like eyes and hands. The second stage, called RefineNet, aims to explicitly handle "hard" keypoints by integrating all levels of feature representations from the GlobalNet with an online hard keypoint mining loss.

All approaches mentioned above are multi-stage. Simple Baseline [58] marks the beginning of a single-stage pipeline. Indeed, the algorithm complexity grows simultaneously with the progress achieved in the human pose estimation task, which makes the analysis of the proposed approaches more difficult. To address this issue, Xiao et al. [58] provided a simple and effective version of the Stacked HourGlass network that adds deconvolutional layers to the backbone network instead of using skip connections.

Most existing frameworks involve a set of top-down bottom-up cascaded sub-networks to switch from high to low, then to high resolution. Sun et al. [59] proposed a novel pipeline, called HighResolution Network (HRNet), that can maintain a high-resolution map all over the model while exploiting the low-resolution maps through parallel branches. This last strategy may reduce the effect of the up-down sampling process. As illustrated in Fig. 11, HRNet is composed of several parallel stages where the first stage presents the highest resolution subnetwork. Exchange units are used to concatenate progressively high-to-low sub-networks.

Multi-stage methods may be more appropriate for the pose estimation task as they naturally provide high spatial resolution and are more flexible. Nevertheless, performances achieved by single-stage methods are still better. Li et al. [60] proved that the low performance of these multi-stage methods is due to the unsuitable choice of the network architectures and that adding few improvements can significantly improve the overall performance. To this end, the proposed Multi-Stage Pose network (MSPN) incorporates three improvements. First, they adopted GlobalNet of CPN [57] as the single-stage module of MSPN. Second, they proposed to aggregate features across different stages to deal with the information that is affected by the repeated down-sampling and up-sampling steps. Additionally, coarse-to-fine supervision is proposed to further boost the system performance.

Unlike most state-of-the-art methods that focus on the network structure to improve the pose estimator performance, Huang et al. [61] highlighted the critical importance of data processing. According to their study, the introduced data processing, which mainly incorporates data transformation and encoding-decoding process, caused a non-aligned between results obtained by flipping common strategy and the original ones in inference, as well as a statistical error in encoding-decoding during both training and inference. To tackle these issues, they proposed a principled Unbiased Data Processing (UDP) which uses a unit length-based measurement (representing intervals between pixels) instead of pixels, and combined classification and regression for encoding-decoding. The proposed UDP can be used in most top-down pose estimators.

Graph convolutional network is another convolutional network category that is used for the 2D pose estimation task. To model the interconnected structure of body keypoints, Bin et al. [62] designed a novel approach named Pose Graph Convolutional Network (PGCN) based on original graph convolutional networks to generate a graph between

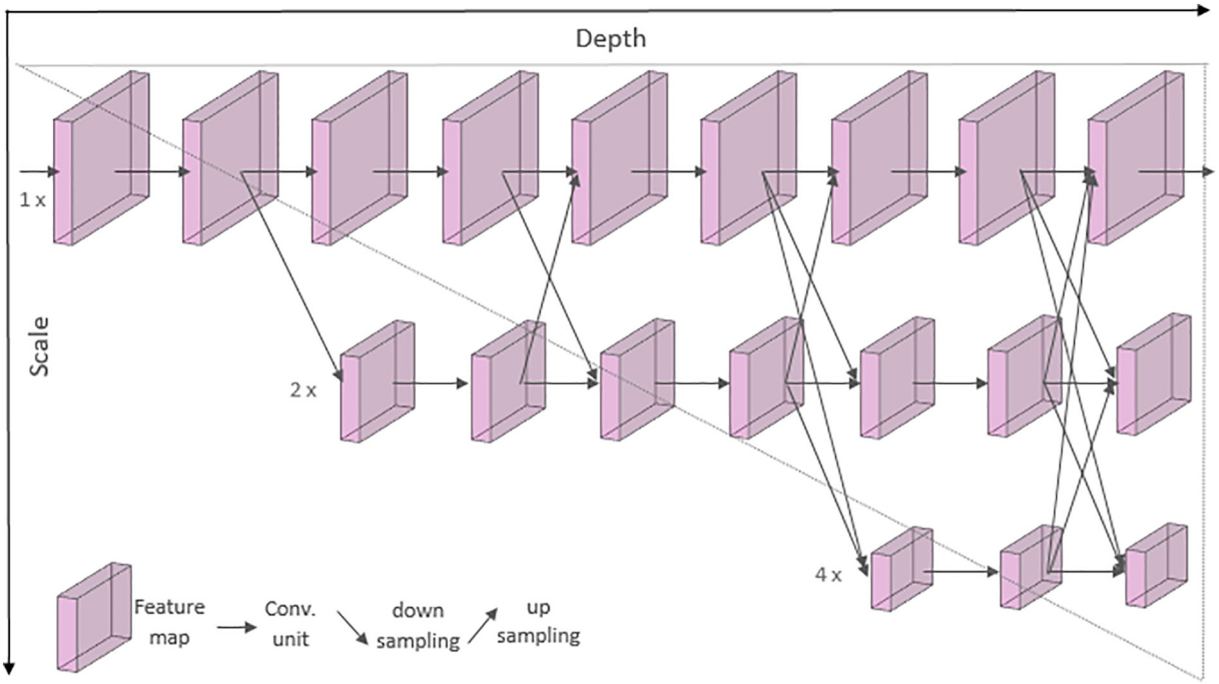

Fig. 11. HRNet architecture.

body keypoints. Each key point is represented by a tensor composed of several feature maps to keep accurate spatial information. Also, PGCN introduced an attention mechanism to capture the structured information between keypoints. Then, it mapped the resulting keypoints graph to a set of structure-aware keypoints representations. The model presented two modules. The first module is Local PGCN which is used to refine the location of key points locally and the second module, named Non-Local PGCN, aims to capture global underlying contextual information.

To address the high computational cost issue while keeping a high system performance, Zhong et al. [62] proposed a low computational-cost deep supervision pyramid network called DSPNet. First, the transposed convolution, which incorporates the decoder module, is replaced by a lightweight up-sampling unit that uses separable transposed convolution, channel-wise attention, and lightweight self-attention. Then, a novel deep supervision pyramid network is used to introduce multi-scale supervision while keeping the same number of parameters.

4.2.2. Bottom-up methods

Most state-of-the-art methods present a top-down architecture. Nevertheless, to overcome all drawbacks related to dealing with each person individually, bottom-up pose estimation pipelines introduced two steps. The first step focuses on detecting and localizing the joints of all persons on the image. The second step consists of associating the detected joints by person instance. This strategy makes bottom-up methods faster and more suitable for real-time applications.

DeepCut [64] jointly estimates the poses of all people present in an image by minimizing a joint objective. The approach used a Fast R-CNN [65] based body part detector to first detect all body part candidates, then introduced a partitioning and labeling formulation to label every body part. To predict the final pose, these parts are grouped with integer linear programming. On the other hand, Deepercut [66] which is an improved version of DeepCut, changes the manually calculated features by introducing an extremely deep part detector based on ResNet [55] to generate body parts proposals. To improve the model performances and to reduce the inference run-time, novel image-conditioned pairwise terms between body parts are introduced to assemble the proposals into a variable number of consistent body part configurations. Also, DeeperCut introduced a novel incremental optimization method that explores the search space more efficiently to reduce the run-time and boost the model accuracy.

Almost real-time multi-person pose estimators present a bottom-up pipeline. Cao et al. [67] proposed a novel algorithm, known as OpenPose. This algorithm uses an explicit nonparametric representation of the keypoints association, called Part Affinity Fields (PAFs), which encodes both position and orientation of human limbs. The network architecture is composed of two branches as shown in Fig. 12. The first branch aims to predict a set of 2D confidence maps S of body part locations and the second branch define a set of 2D vector field of PAFs. Experiments proved the efficiency of the proposed approach to detect accurate poses, as well as its robustness against a high number of people in the input image.

Osokin [68] proposed an improvement version of OpenPose. To use it in edge devices, the study proposed a simpler architecture that reduces the computational complexity in real-time applications. To this end, they proposed a lightweight network backbone that replaced VGG [69] by MobileNet V1 [70]. They also introduce dilated convolutions and use a single prediction branch in the initial stage and the refinement stage. Finally, they refined the code and removed extra memory allocations, and parallelized keypoints extraction to make the code faster. Inspired by OpenPose, Kreiss et al. [71] proposed a new approach that addresses challenges related to low-resolution and partially occluded people using the Part Association Field (PAF). The network refers to two kinds of maps and presents two blocks. The first block is designed to predict the confidence score, the precise localization, and the size of each body part or joint, which they called Part Intensity Field (PIF). The second block predicts the PAF map, the association between parts. They also use the Laplace loss for regressions. The overall model, illustrated in Fig. 13, is called PifPaf, as it incorporates the two maps PIF and PAF.

Bottom-up approaches have to handle large variations in human scale. Using multi-scale pyramid networks and larger input size may address this issue and produce more accurate pose estimation heatmaps. However, these techniques suffer from a very high computational complexity. Inspired by HRNet [59], Cheng et al. [72] proposed a bottom-up approach, named Higher Resolution Network (HigherHRNet), that

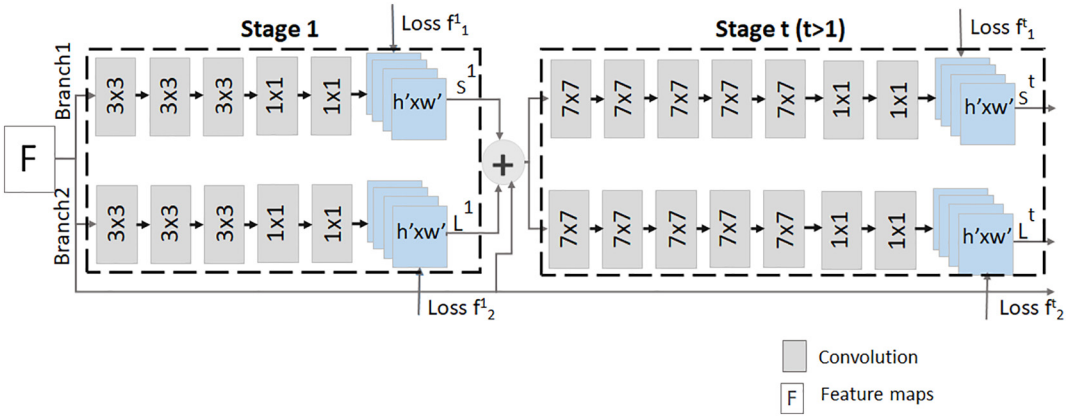

Fig. 12. OpenPose architecture.

addresses this issue while generating a high multi-resolution heatmap. Indeed, HigherHRNet is an extended version of HRNet with additional deconvolutional modules. From another perspective and inspired by HRNet architecture, Cheng et al. [73] proposed a novel one-shot learning network, named ScaleNAS, to explore different depths for multi-scale feature representations. The network includes multi-scale aggregation search space that can address multiple vision recognition tasks. Derived from ScaleNAS, authors proposed ScaleNet-P as a bottom-up pose estimation model that surpassed HigherHRNet performances. Another top-down ScaleNet-P version is also proposed. In addition to that, multi-resolution supervision is introduced during training. During the keypoint grouping process, keypoints are clustered based on the L2 distance. To adjust the standard deviation for each keypoint based on its scale and uncertainty, Luo et al. [74] proposed a Scale-Adaptive Heatmap Regression (SAHR) framework. First, Gaussian kernels with the same standard deviation are used for all keypoints. Then, scale maps with the same shape of ground truth heatmaps are predicted. Finally, the standard deviation for each keypoint is computed using a pointwise operation. To handle the imbalance between fore-background samples, a Weight-Adaptive Heatmap Regression (WAHR) is implemented to focus on hard samples. Artacho et al. [75] proposed OmniPose, an end-to-end single pass framework. To incorporate contextual information, the network includes an improved version of HRNet to leverage multi-scale feature representations. An advanced waterfall decoder module WASPv2 [49] is also introduced to generate a heatmap and a confidence map for each joint. The network does not require any postprocessing. To decrease the number of parameters and the computational cost of the model, authors proposed the lightweight OmniPose-Lite architecture.

From another perspective, Geng et al. [76] emphasized the fact that focusing on the keypoint regions can help to accurately regress the joint localizations. They proposed Disentanged Keypoint Regression (DEKR) which includes adaptive convolutions and multi-branch structure to produce a disentangled representation for each keypoint region.

4.3. Results and discussion

As mentioned earlier, several frameworks have been proposed to predict the 2D human joints locations. With the introduction of deep learning techniques, these frameworks have achieved very high performances. Based on the overall architecture, we have classified the studies into different categories. Each category has its strengths and weaknesses.

Regression-based models have achieved significant performances due to their ability to learn non-linear feature representations. Based on the prior works, these methods are fast, they can be trained by an end-to-end model, they generate a continuous output, and more particularly, they could be extended to 3D scenarios without major adjustments. Nevertheless, the major drawback of these approaches is that they are not suitable for multi-pose scenarios. Moreover, even though joint localization is essentially a regression problem, deep ConvNets may face difficulties in regressing accurate heatmaps for body keypoints with high occlusions, as well as for background that seems like body parts. Consequently, detection-based methods achieved state-of-the-art accuracy on 2D pose estimation. Heatmaps can provide richer supervision information by capturing the spatial information related to the keypoint location. Nevertheless, these methods apply a post-processing step to transform heatmaps to joint location, usually by using the non-differentiable argmax function, which prevents the training from being end-to-end. Moreover, the use of high-resolution heatmaps increases the system efficiency, but it is highly demanding in terms of computation and memory. Lastly, such methods could be hardly adapted to 3D

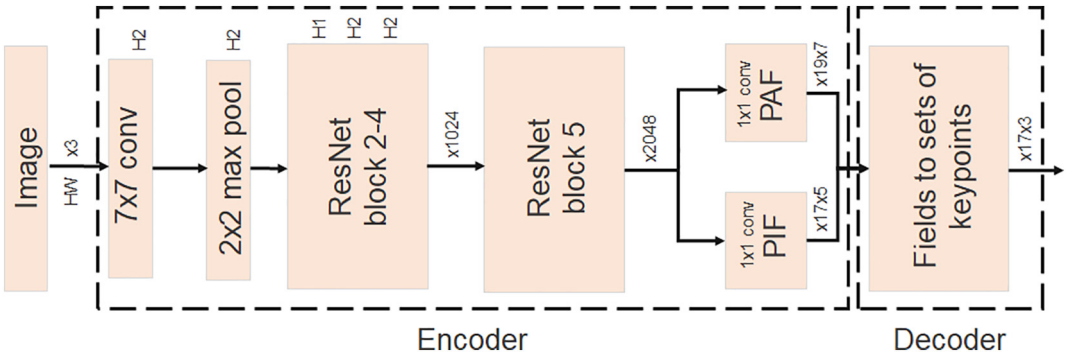

Fig. 13. PifPaf architecture.

scenarios. We summarize the 2D single-person approaches on the MPII dataset and the LSP dataset respectively in Tables 1 and 2.

Similarly, multi-person pose estimation faces several challenges. These challenges are mainly related to the fact that the position, the scale, and the number of persons in the image are unknown. Also, the interaction and the overlap between persons affect the visibility of joints. Many other factors need to be addressed. Top-down approaches are widely used. First, they take advantage of existing single-person pose estimators. Besides, they are generally less sensitive to person-scale variances, as they can normalize all cropped person regions to approximately the same resolution. However, as they involve a person-detector in the first stage, they particularly suffer when people are obscured by other people or are close to each other. The joint localization process fails if the person detector does not give an accurate result. Moreover, this strategy is not efficient in crowd scenes, mostly in real-time applications where the execution time is critical. Furthermore, the computational time is highly related to the number of persons in the image as every person pose has to be estimated separately. Also, this criterion depends on the employed person-detector. There is a trade-off between the accuracy and the computational complexity of this component. We summarize top-down results on MPII dataset in Table 3.

Bottom-up approaches have shown high robustness towards all top-down pipeline issues by providing models where the run-time complexity is detached from the number of persons in the image. However, in terms of performances, there is still a wide difference between bottom-up and top-down methods. First, the scale variations of people in the image can significantly degrade the performance of inferring the pose, since the network is not able to generate accurate heatmaps. Besides, in most cases, they fail to predict an accurate pose for smaller persons. Moreover, the bottom-up approaches face some difficulties associating keypoints by person instance in occluded scenes. A comparison of multi-person approaches on COCO test-dev set is provided in Table 4.

In summary, it is difficult to conclude which method is better since both top-down and bottom-up methods are widely used in recent works with the introduction of various schemes to reduce their drawbacks.

Another classical challenge that can limit the method performance is the availability of rare poses. Current datasets are large, however, they contain limited training samples for unusual poses which may cause model bias. Applying data augmentation techniques can help with this issue.

Table 1
Comparison of previous 2D single-person approaches on MPII test set (PCK@0.5).

2D single-person pose estimation								
Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Regression-based approaches								
Carreira et al. [34]	95.70	91.60	81.50	72.40	82.70	73.10	66.40	–
Sun et al. [35]	97.50	94.30	87.00	81.20	86.50	78.50	75.40	86.40
Luvison et al. [36]	98.10	96.60	92.00	87.50	90.60	88.00	82.70	91.20
Zhang et al. [37]	97.20	95.90	91.20	86.70	89.70	86.70	84.00	90.60
Detection-based approaches								
Sun et al. [39]	98.10	96.20	91.20	87.20	89.80	87.40	84.10	91.00
Yang et al. [41]	98.50	96.70	92.50	88.70	91.10	88.60	86.00	92.00
Chou et al. [43]	98.20	96.80	92.20	88.00	91.30	89.10	84.90	91.80
Chen et al. [44]	98.60	96.40	92.40	88.60	91.50	88.60	85.70	92.10
Nie et al. [46]	98.60	96.90	93.00	89.10	91.70	89.00	86.20	92.40
Ke et al. [40]	98.50	96.80	92.70	88.40	90.60	89.30	86.30	92.10
Su et al. [47]	98.70	97.50	94.30	90.70	93.40	92.20	88.40	93.90
Bulat et al. [48]	98.80	97.50	94.40	91.20	93.20	92.20	89.30	94.10
Artacho et Savakis [49]	–	–	–	–	–	–	–	92.70

Table 2
Comparison of previous 2D single-person approaches on LSP test set (PCK@0.2).

2D single-person pose estimation								
Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Regression-based approaches								
Luvison et al. [36]	97.50	93.30	87.60	84.60	92.80	92.00	90.00	91.10
Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Detection-based approaches								
Sun et al. [39]	94.90	88.80	77.60	70.70	88.90	84.80	80.50	83.70
Yang et al. [41]	98.30	94.50	92.20	88.90	94.40	95.00	93.70	93.90
Chou et al. [43]	98.20	94.90	92.20	89.50	94.20	95.00	94.10	94.00
Chen et al. [44]	98.50	94.00	89.80	87.50	93.90	94.10	93.00	93.10
Bulat et al. [48]	98.70	95.70	93.10	90.30	95.80	95.60	94.80	94.80
Artacho et Savakis [49]	–	–	–	–	–	–	–	94.50

Table 3
Comparison of previous 2D multi-person approaches on MPII test set (Top-down approaches are evaluated based on PCK@0.5 and Bottom-up approaches are evaluated based on AP).

2D multi-person pose estimation								
Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Top-down approaches on MPII single-person test set								
Wei et al. [53]	97.80	95.00	88.70	84.00	88.40	82.80	79.40	88.50
Newell et al. [38]	98.20	96.30	91.20	87.10	90.10	87.40	83.60	90.90
Fang et al. [56]	91.30	90.50	84.00	76.40	80.30	79.90	72.40	82.10
Sun et al. [59]	98.60	96.90	92.80	89.00	91.50	89.00	85.70	92.30
Li et al. [60]	98.40	97.10	93.20	89.20	92.00	90.10	85.50	92.60
Bin et al. [62]	98.00	96.90	92.70	89.00	91.80	89.40	86.10	92.40
Zhong et al. [63]	–	–	–	–	–	–	–	92.50
Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
Bottom-up approaches on MPII multi-person test set								
Pishchulin et al. [64]	73.40	71.80	57.90	39.90	56.70	44.00	32.00	54.10
Insafutdinov et al. [66]	78.40	72.50	60.20	51.00	57.20	52.00	45.40	59.50
Cao et al. [67]	91.20	87.60	77.70	66.80	75.40	68.90	61.70	75.60

5. 3D human pose estimation

The 3D human pose estimation task consists of producing a three-dimensional output that indicates the spatial position of the person articulations. It serves as a fundamental key for several applications in computer vision including human-computer interaction, 3D augmented reality, gaming, activity recognition, and so on. With the introduction of deep convolutional neural networks, this field has achieved significant advancement in recent years.

Similar to 2D, 3D approaches are classified into two main groups: single-person approaches which present vast literature, and multi-person pose approaches that are mostly unexplored due to the non-availability of annotated datasets. Typically, most real-time applications require multi-person processing. Despite its interesting potential, the use of single-person methods has a major disadvantage since the processing time is increased by the number of people in the image, which makes it unsuitable for the analysis of crowded scenes. In the following subsections, we detail recent approaches used for 3D pose estimation.

5.1. Single-person pipeline

For the single-person 3D human pose estimation task, prior work can be divided into two main categories. The first category consists of one-stage approaches that regress the 3D poses directly from a given image. The second category uses two separate stages where firstly a

Table 4
Comparison of previous 2D multi-person approaches on COCO test-dev set.

2D multi-person pose estimation										
Method	AP	AP.5	AP.75	AP(M)	AP(L)	AR	AR.5	AR.75	AR(M)	AR(L)
Top-down approaches										
Papandreou et al. [54]	68.50	87.10	75.50	65.80	73.30	73.30	90.10	79.50	68.10	80.40
Fang et al. [56]	72.30	89.20	79.10	68.00	78.60	–	–	–	–	–
Chen et al. [57]	73.00	91.70	80.90	69.50	78.10	79.00	95.10	85.90	74.80	84.70
Xiao et al. [58]	73.70	91.90	81.10	70.30	80.00	79.00	–	–	–	–
Sun et al. [59]	74.90	92.50	82.80	71.30	80.90	80.10	–	–	–	–
Li et al. [60]	78.10	94.10	85.90	74.50	83.30	83.10	96.70	89.80	79.30	88.20
Huang et al. [61]	76.50	92.70	84.00	73.00	82.40	81.60	–	–	–	–
Zhong et al. [63]	73.70	–	–	–	–	79.10	–	–	–	–
Bottom-up approaches										
Cao et al. [67]	61.80	84.90	67.50	57.10	68.20	–	–	–	–	–
Kreiss et al. [71]	50.00	73.50	52.90	35.90	69.70	55.00	76.00	57.90	39.40	76.40
Cheng et al. [72]	70.50	89.30	77.20	66.60	75.80	–	–	–	–	–
Cheng et al. [73]	71.60	–	–	–	–	–	–	–	–	–
Luo et al. [74]	72.00	90.70	78.80	67.80	77.70	–	–	–	–	–
Artacho et al. [75]	76.40	92.60	83.70	72.60	82.60	81.20	–	–	–	–
Geng et al. [76]	70.00	89.40	77.30	65.70	76.90	75.40	–	–	69.70	83.20

2D pose is estimated and then the resulting joints are projected to 3D space. Considering the quantity and the quality of the labeled 3D data, the last category may reduce the overfitting effect. However, lifting the pose from 2D to 3D may lead to the loss of depth information.

5.1.1. One-stage approaches

In this section, we present the 3D approaches that directly regress the 3D pose without the need to go through a 2D pose estimator.

Li and Chan [77] are the first to show that deep neural networks can achieve a reasonable accuracy in inferring 3D human pose from a given image. The proposed network is trained using two techniques. The first technique is a heterogeneous multitask framework that jointly trains pose regression tasks with multiple body part detection tasks. The second technique is a pre-training strategy where the pose regressor is initialized using a network trained for body part detection, and then refine the network using only the pose regression task. The purpose of the regression task is to predict the positions of the articulation points relative to the root joint position. For the body detection tasks, each of them is used to classify whether a local window contains the specific joint. Indeed, body joints are highly correlated. Through this study, it has been proven that a deep network can learn on its own way the dependencies between different body parts without any explicit intervention.

Usually, techniques that use ConvNets to directly regress the 3D pose involve a kind of ignorance of the dependencies between human joints. To address this problem, while maintaining a low computational cost at inference time, Tekin et al. [78] proposed a deep learning regression framework for structured 3D human pose prediction from monocular images. The last approach relies on an overcomplete auto-encoder that projects body joint positions to learn a high-dimensional latent pose representation. After training the auto-encoder, they train a CNN-based network and map it to the resulting high-dimensional pose representation. As a consequence, this may impose implicit constraints on human posture, maintain body statistics, and increase the prediction accuracy. Finally, they connect the decoding layers of the auto-encoder to the CNN network and fine-tune the whole model. The overall architecture is shown in Fig. 14. To enhance well the method performances by exploiting the joint dependencies, Sun et al. [35] proposed a structure-aware regression method that replaces the joint representation with bones. This method aims to infer both 2D and 3D poses.

To address the detection-based pipeline limitations, the work proposed in [79] highlighted two important contributions. First, authors considered the pose estimation task as a keypoint localization problem in a discretized 3D space. That is why they trained a ConvNet to predict the likelihood per voxel for each joint. The obtained volumetric representation is much more sensitive to the 3D nature of this task. To further improve the initial estimation, they used a coarse-to-fine prediction scheme that increases progressively the supervision volume resolution for the zth-dimension. This step addresses the large dimensionality increase that takes place more and more in the 3D domain and enables iterative refinement. In the study made by Sun et al. [80], the proposed unified approach, named integral regression, replaces heatmaps with

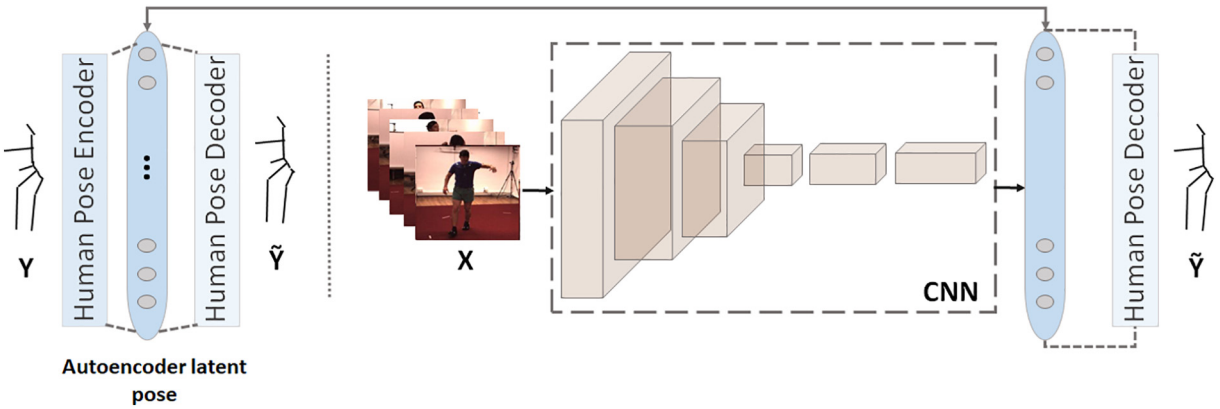

Fig. 14. 3D pose estimator based on an auto-encoder.

joint location coordinates. The joint location is obtained as the integration of all locations in the heatmap, weighted by their probabilities (normalized from the probabilities). In such a manner, the merits of both approaches are combined, while addressing their limitations.

Due to the difficulty of obtaining the 3D ground truth of the human pose in outdoor environments, most existing datasets are labeled in a laboratory environment using motion capture systems. Consequently, this can cause overfitting during training hence the variations in background, viewpoints, and lighting are limited in the indoor environment. Yang et al. [81] proposed an adversarial learning paradigm to transform the 3D human pose structures learned from annotated constrained indoor 3D pose datasets into unconstrained outdoor environments where annotations are non-available. Following the GANs (Generative Adversarial Networks) structure, the network is composed of a generator and a discriminator. The generator is a 3D pose estimator that generates samples in a manner that confuses the discriminator, which in its turn attempts to differentiate fake samples from real samples. The following helps to enhance the pose estimator to generate valid poses with indoor images, as well as outdoor images. The discriminator accuracy has a direct influence on the pose estimator. For this reason, a new multi-source geometric discriminator has been included to reduce the gap between the predicted poses and the ground truth poses. Weakly supervised methods are also recommended since they can use non-labeled 2D poses or multi-view images that require less supervision. Nevertheless, they still require a large dataset to achieve good performance while avoiding overfitting. To learn a geometry-aware body representation from multi-view images without annotations, Rhodin et al. [82] suggested exploiting images of the same person taken from multiple views to learn a latent representation that captures the 3D geometry of the human body. To achieve this purpose, they proposed to use an encoder-decoder that aims to generate multi-view images from a single image. Thus, mapping to the 3D pose by introducing the latent representation in supervised learning is much simpler, and using fewer examples during the training process can be sufficient.

5.1.2. Two-stage approaches

In this section, we describe approaches that address the 3D pose estimation task based on the 2D joint locations. The basic architecture of these models implements cascaded 2D and 3D regressors.

Real-time application approaches require a trade-off between a high degree of accuracy and negligible inference time. On the other hand, pre-processing steps such as the extraction of bounding boxes make the process quite difficult to set up. According to these facts, Mehta et al. [83] proposed the first real-time method to capture the global skeleton of the 3D human poses, by combining a CNN pose regressor with kinematic skeleton fitting. Thus, they eliminate the need to extract bounding boxes. The novel pose formulation regresses jointly the 2D and the 3D joint positions. Then, a real-time kinematic skeleton fitting method uses the CNN output to produce 3D global pose reconstructions based on a coherent kinematic skeleton. To extend the 2D heatmap formulation to the 3D, they use location maps.

To integrate rich spatio-temporal joint dependencies, Lin et al. [84] proposed a data-driven approach that includes 2D spatial relationship, 3D geometry, and temporal smoothness using the Recurrent 3D Pose Sequence Machine (RPSM) network. RPSM captures long-term dependencies across several body parts for 3D pose prediction and enhances temporal consistency between the predictions of sequential frames. Specifically, it recursively refines predicted 3D pose sequences by detecting what has already been learned previously. Each RPSM stage consists of a 2D pose module that extracts image-dependent pose representations, a feature adaptation module that transforms the representation from 2D to 3D domain, and a recursive 3D pose module that regresses the 3D poses. These three modules are then assembled in a sequential prediction framework to refine the predicted poses with recurrent stages.

As a solution to the lack of large-scale 3D pose annotation datasets, several works take advantage of the large-scale 2D datasets either by using a weakly-supervised paradigm or a self-supervised paradigm. Zhou et al. [85] proposed a weakly-supervised transfer learning method that uses mixed 2D and 3D data to train a deep convolutional neural network. This approach uses the 2D pose annotations of the wild images as weak labels to infer 3D poses. The network includes a 2D estimator and a 3D depth regressor. Both modules are connected with the intermediate layers of the 2D module. Thus, common features between 2D and 3D pose tasks can be shared easily. To integrate both fully-labeled and weakly-labeled data, a new loss function is induced based on a geometric constraint. The overall architecture is shown in Fig. 15.

Dabral et al. [86] suggested two novel anatomically inspired loss functions, respectively named as illegal-angle and symmetry loss, and use them with a weakly supervised learning framework to train the model using both large-scale outdoor 2D data and limited indoor 3D data. The proposed loss function is applied to the 2D images during training and ensures that the predicted 3D pose does not violate anatomical constraints, such as left–right human body symmetry. Thom et al. [87] proposed a novel method that infers jointly the 2D and the 3D pose. The approach combines the probabilistic aspect of the 3D human pose with a multi-stage CNN architecture and uses the ground truth of the 3D pose to refine the 2D joint localization. Features captured by the 3D human pose model are then integrated into the CNN architecture as an additional layer that transforms 2D joint coordinates into 3D. The proposed architecture can be trained using only 2D annotations. A novel self-supervised learning method [88], called Epipolar, is proposed to estimate the 3D human pose. Epipolar does not require any 3D labeled data and presents two branches that start with the same pose estimation network. In the first branch, EpipolarPose estimates 2D poses from multi-view images and then uses the epipolar geometry and the camera geometry to get the 3D pose ground truth. The pose estimator layers are frozen during training. In the second branch, a soft argmax function is used to transform 2D heatmaps to 3D pose. EpipolarPose works with an arbitrary number of cameras (at least 2) and does not need 3D supervision or extrinsic camera parameters, but it can use them if they are provided. Gong et al. [89] suggested a novel online data augmentation algorithm, called PoseAug. The network includes three modules to generate new data: a differentiable augmentor module that applies three types of geometry augmentation, a 2D-3D pose estimator, and a discriminator that ensures the plausibility of the augmented data based on a novel part-aware KCS [90] representation.

In previous work, it is not easy to assume whether the prediction error is caused by the 2D pose estimation stage, or by projecting 2D poses to 3D. To this end, Martinez et al. [91] proposed a framework that uses 2D ground truth poses as an input. They proved that a cascaded fully connected layer network with residual connections can be efficient. Other works address the lifting module to enhance the model performance. In the study of Ci et al. [92], the proposed approach combines the advantages of both GCN [93] and FCN [80,91], while addressing their drawbacks. The network includes the state-of-the-art HourGlass network [38], followed by the proposed LCN network that aims to lift the 2D pose to the 3D domain. LCN uses fully connected layers with residual links. Compared to the FCN, each joint of the LCN is only connected to its related joints according to human anatomy. Thus, the number of learnable parameters is significantly reduced. The overall architecture is presented in Fig. 16. Unlike the GCN, it excludes the weight-sharing scheme and assigns dedicated filters to each joint which cures the representation capacity.

To estimate 3D human pose from videos, Pavllo et al. [94] integrated a supervised module with an unsupervised one that acts as a regularization. First, a CNN network with residual connections is used to transform a given sequence of 2D poses through temporal convolutions. The convolutional module ensures parallelization on both the batch and the time dimension. Dilated convolutions are also used to model long-term dependencies as well as retaining efficiency. Second, they

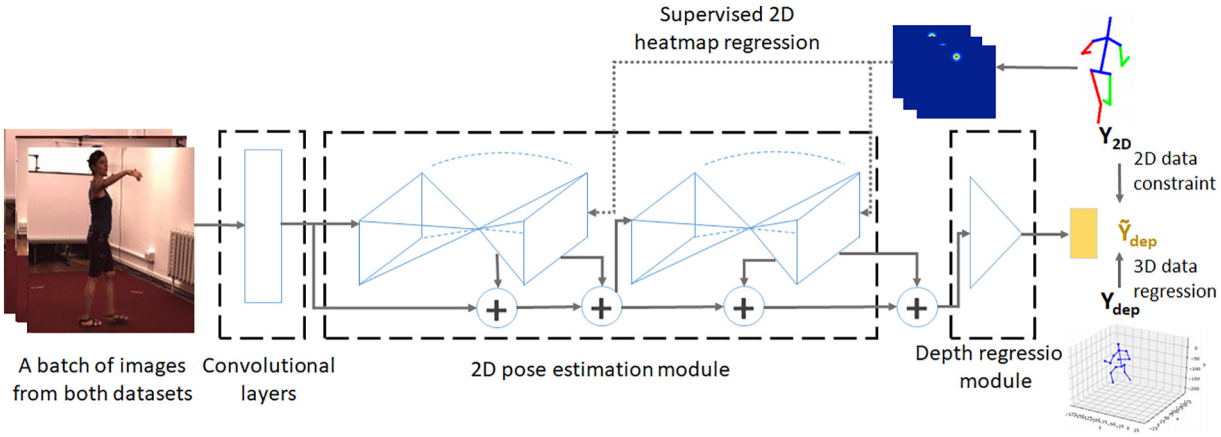

Fig. 15. Weakly-supervised 3D pose estimator.

introduce a semi-supervised approach that provides back-projection by using unlabeled video data. Such an approach improves accuracy in settings where the availability of labeled 3D data is limited. To train the model, this method uses unlabeled video and needs only intrinsic camera parameters, this makes it very useful in various scenarios where the motion capture is a challenging task.

Estimating human pose with multi-view images can be very interesting. First, it allows getting the ground truth in the outdoor environment. Motion capture systems usually suffer from certain weaknesses like the failure to capture rich pose representations (e.g., to estimate hand and face poses next to limb poses). Also, it can be used directly to follow the human pose in real-time, since multi-camera installations become progressively available for various applications, such as sports, etc. Takahashi et al. [95] proposed a new algorithm to estimate the human 3D pose from multi-view videos captured by unsynchronized and non-calibrated cameras. First, the network detects 2D joint positions from multi-view videos using a 2D pose detector. To avoid detection errors, the suggested method applies a median filter after applying a cubic spline interpolation method to the output data. Then, two cameras are selected where their parameters are initialized by the standard SfM (Structure from Motion) approach and decomposed into extrinsic parameters to estimate the 3D joints through triangulation. In [96], two novel solutions for multi-view 3D human pose estimation are suggested. The proposed study is based on a new learnable triangulation method that combines 3D information from several 2D views to reduce the number of views needed for an accurate estimation of the human 3D pose. The first technique is an algebraic triangulation with the addition of confidence weights estimated from the input images. The second solution used a new volumetric method based on the dense geometric aggregation of information from different views that aims to model a human pose prior. The aggregated volume is then refined using convolutions to produce the final 3D heatmaps. The suggested system is robust to occlusions and partial views and it explicitly uses the camera settings as an independent input. However,

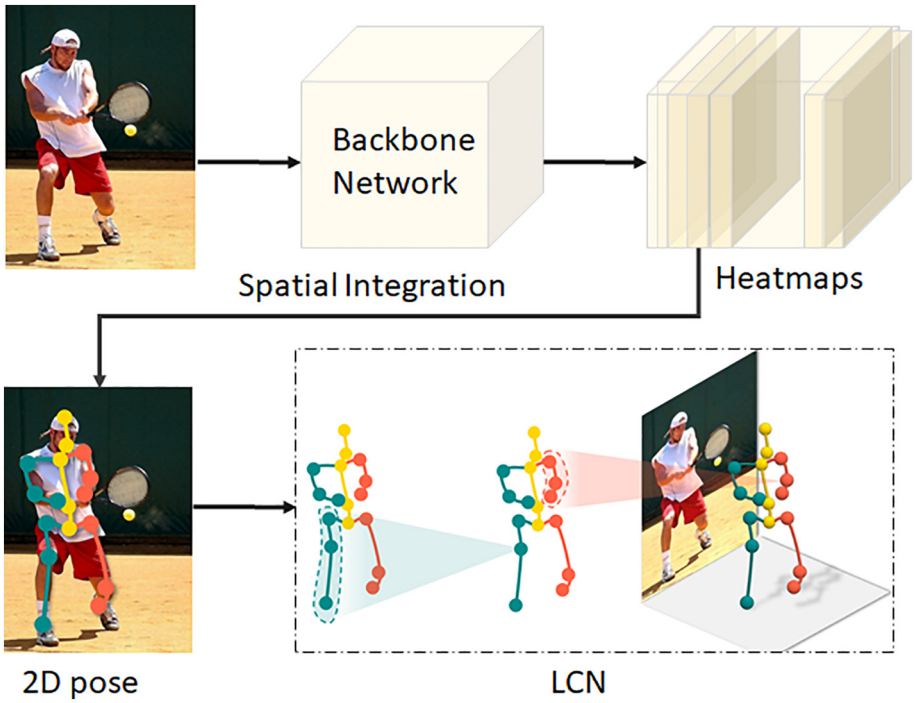

Fig. 16. 3D pose estimator using LCN.

since the triangulation approach relies on predicting algebraic triangulation, it is consequently essential to provide at least two camera views that observe the pelvis, which could be problematic for certain applications. A lightweight solution [93] is also proposed to estimate 3D human pose from multi-view images captured with calibrated cameras. First, the suggested approach uses 3D geometry to combine input images with a unified latent camera-independent pose representation. Then, the lightweight network predicts 2D joint locations based on the resulting representation. Finally, the 2D predicted joints are lifted to 3D using a differentiable Direct Linear Transform (DLT) layer. Wandt et al. [98] proposed a self-supervised framework, named CanonPose that leverages unlabelled multi-view data to infer 3D poses. The network involves multiple neural networks where each sub-network produces a 3D pose in a canonical rotation. Outputs across different views are combined to produce the final pose.

To take advantage of the complex dependencies and correlation information between different parts, Wang et al. [99] proposed Depth Ranking 3D Human Pose Estimator (DRPose3D), to estimate 3D pose based on 2D joint locations and a ranking matrix that illustrates the depth relationships between each pair of human joints. In this way, the 3D pose estimation task turns to a series of classification problems that can be overcome by deep neural networks. The ranking matrix is generated using a network called "Pairwise Ranking Convolutional Neural Network" (PRCNN). After defining the depth ranking, the generated matrix is used with the 2D joint locations to infer 3D poses. The obtained results show that depth ranking is an important geometrical knowledge that provides relevant information regarding the 3D pose estimation task. Liu et al. [100] suggested a new LSTD (Long Short-term Dependency-aware) module incorporated into a CNN architecture to reinforce intermediate convolutional feature maps for the 3D pose estimation task. To this end, LSTD module is implemented based on the Graphical ConvLSTM (Long Short-term Dependency-aware). To refine the designed LSTD module, they added a software modulator, the context consistency gate (CCG), which estimates the consistency of the convolutional features with their contextual information and adjusts these features appropriately to reinforce them. The final architecture consists of several convolutional layers and LSTD modules cascaded together to build a deep feature boosting network.

2D detection systems are mostly noisy and not very reliable because of motion blur and self-occlusion within video streams. Prior work adopts simple structural constraints to facilitate joint 3D predictions. However, such processing is not enough to achieve significant improvements for this task. Moreover, most of the current approaches formulate this task as a regression problem. This strategy does not take the inherent kinematic structure of the human subject into consideration which decreases the system performances. To this end, Xu et al. [101] proposed an approach that systematically incorporates kinematic analysis into deep CNN for efficient use of prior human knowledge. They first refine the 2D inputs to filter noise. Also, they design a new optimization scheme for 2D keypoints under the constraint of perspective projection, to facilitate the kinematic structure correction of the noisy 2D inputs. The next step consists of decomposing the articulated motion based on the rigid body assumption. Specifically, the 3D regression task is divided into two complementary sub-tasks presenting the length and the direction estimation. Lastly, they exclude the non-reliable joints from predictions and consider only the reliable parts.

5.2. Multi-person pipelines

As mentioned earlier, multi-person pipelines are more suitable for real-world applications. However, research in the 3D multi-person pose estimation field is still limited. We mention in the next subsection some recently proposed approaches.

5.2.1. Top-down methods

Similar to the 2D domain, top-down methods use a person detector to crop regions that contain one person and then perform a 3D single-person pose estimator. In the work realized by Rogez et al. [102], a novel end-to-end localization-classification-regression framework, called LCR-Net, is proposed. The network architecture consists of a pose proposal generator that provides a list of candidate poses, a classifier that generates a score for each pose proposal, and a regressor that aims to refine pose proposals in 2D and 3D. Final poses are obtained by combining the resulting pose proposals. Moon et al. [103] suggested a camera distance-aware framework that can be smoothly integrated with most 3D human detection and human pose estimation systems. The framework architecture is composed of DetectNet, a human detector, RootNet, a 3D human root absolute localization network that predicts the camera-centered coordinates of detected human roots, and PoseNet, a 3D single-person root-relative pose estimation module that provides 3D root-relative pose estimation for each detected human.

Dong et al. [104] introduced a novel fast and robust method that is based on a multi-way matching algorithm that aims to cluster the detected 2D poses among multi-views. Moreover, they combined geometric and appearance information for cross-view matching. Then, the 3D pose is predicted for each individual separately from the resulting matched 2D poses.

As mentioned before, heatmap representations have multiple limitations, especially when we look for inferring multi-person poses. PandaNet (Pose estimAtioN and Detection Anchor-based Network) [105] is a new single-shot, anchor-based, and multi-person 3D pose estimation approach. The network detects a bounding box for each person in a given image, and then fits every proposal region into a regression of 2D and 3D poses in a single forward pass. To store the full 3D pose of a given subject, the model replaces heatmaps with an anchor-based representation. Due to this formulation, a single output pixel is enough to store the entire subject pose. Besides, a "Pose-Aware" anchor selection strategy is introduced to address people overlapping. Furthermore, people in images have different scales, which leads to uncertainties in the joint coordinates. Thus, the study introduces a method to automatically optimize the weights associated with different people scales and joints.

The multi-view 3D pose estimation task has two major association issues. First, it requires associating the joints of the same person by either top-down or bottom-up approaches. Then, it involves the association of the 2D poses of the same person with different views which are not stable when there are occlusions. VoxelPose [106] is a multi-person 3D pose estimator that works directly in 3D space by collecting information from all camera views. The first step consists of estimating 2D heatmaps for each view to encode the per-pixel likelihood of all joints. Unlike previous work, VoxelPose projects the heatmaps of all views into a common 3D space to obtain a feature set that provides accurate estimates of the 3D positions of all joints. A Cuboid Proposal Network (CPN) is used to locate all people in the scene by predicting several 3D cuboid proposals from the 3D feature volume. Then, for each proposition, a volume of characteristics is fed into a Pose Regression Network (PRN) to estimate the 3D pose.

Unlike most previous multi-person pose estimation studies that require creating or simulating a 3D human pose dataset for the training phase, Dabral et al. [107] proposed a novel method that relies only on 2D pose datasets. The proposed architecture, named HG-RCNN is based on Faster-RCNN [50] where every proposal region passed through Hourglass which is used to regress 2D heatmaps. Then, a 2D-to-3D fitting module is used to regress the root-relative 3D poses. As a final step, the resulting keypoints are converted to camera coordinates without using geometric optimization.

To overcome the lack of global consistency of most top-down methods, Wang et al. [108] proposed a new supervision strategy, named Hierarchical Multi-person Ordinal Relations (HMOR). HMOR

incorporates both depth and angle relationships by dividing the interaction information into three levels: human instances, body parts, and joints. For this purpose, an integrated model is employed to perform simultaneously as a human detector, a pose estimator, and a human depth estimator. A coarse-to-fine architecture is also introduced to improve the accuracy of the depth estimator. Epipolar constraints are a fundamental key in previous multi-person multi-camera 3D human pose estimation methods. However, they are not robust in crowd scenes. To this end, Chen et al. [109] adopted a novel multi-view geometry method that is more suitable for this type of scene. The network architecture is composed of a graph model that ensures a fast cross-view matching based on feet assignment across multi-views, and a maximum a posteriori (MAP) estimator that is used to reconstruct the final 3D human poses.

5.2.2. Bottom-up methods

Unlike top-down pipelines, bottom-up approaches first generate all body joint locations, then cluster body parts to each person instance. Since top-down methods use a person detector to infer each human pose separately, they are highly demanding in terms of inference run-time, memory consumption, and computation cost.

To overcome the inference run-time issue, a new single-shot CNN-based method [28] is proposed to jointly predict multi-person poses from a given image for both domain 2D and 3D in a single forward pass. This approach does not require any body extractor to define region proposals. A novel Occlusion-Robust Pose-Map (ORPM) formulation is introduced to predict full-body pose inference even under extreme conditions as string occlusions, by creating redundancy in the encoding. Thus, the proposed method ensures a fixed size of the output regardless of the number of people in the scene. To train this approach, a new multi-person 3D pose dataset, called MuCo-3DHP, is introduced. Regardless the accuracy, the work proposed by Mehta et al. [110] took into consideration the same issues. First 2D and 3D pose features are extracted for all visible joints. Then, another neural network uses these features to generate the completed 3D pose. Finally, the predicted poses are refined using a space–time skeletal model. Using low-resolution representation can also be a solution. However, this can affect the model performances. To this end, Fabbri et al. [111] presented a novel compression-based approach that is composed of three main modules. First, a volumetric heatmap auto-encoder is used to compress the ground-truth heatmaps. Then, a Code Predictor is trained to predict 3D joint locations, which are then fitted into a decoder to get the original representation.

It is not easy to capture the 3D multi-person pose in an unconstrained environment. To tackle this challenge, Kundu et al. [112] proposed a novel unsupervised single-shot framework that adopts a novel neural representation of multi-person 3D pose by unifying the person instance positions with their corresponding 3D pose representation. Unlike most bottom-up approaches, a generative pose embedding is used to avoid the utilization of keypoints grouping operation. The learning process is formulated as a cross-modal alignment problem. Training objectives are also introduced to realize a shared latent space between two different data-flow pathways. The following approach made a remarkable gain in performance while achieving an optimal computational cost.

The absolute depth of human bodies in an image shows occlusions between them which can be very useful in the pose estimation process. Besides, it can reflect the spatial extent of each person in 2D. Zhen et al. [113] proposed a novel single-shot depth-aware approach to better exploit the inter-person depth relationship. The implemented network regresses the root depth map, the 2D keypoint heatmaps, and part affinity fields (PAFs) which are used to group keypoints into human instances.

5.3. Results and discussion

3D human pose estimation is a very critical task in the computer vision field since it can serve as a fundamental tool for several practical applications. Based on the different articles presented below, we summarize and discuss all results in this section.

To measure the 3D pose estimation methods performance on the Human3.6 M dataset, several evaluation protocols have been followed. Most studies adopt Protocol 1 to evaluate performances on the Human3.6 M dataset. For single-person pipelines, we divided the approaches into two categories. The first category includes approaches that directly regress the 3D pose without going through intermediate stages. Thus, we summarize the one-stage approach results in Table 5. The main drawbacks of these approaches are the requirement of a large-scale dataset to avoid overfitting during the training process.

To overcome these drawbacks, two-step approaches are proposed. They split the estimation task into two parts. Firstly, a 2D pose estimator is used to locate joints. The resulting information is then used to lift to 3D pose. This strategy provides the possibility of leveraging the diversity of the 2D datasets. We summarize the obtained results of these approaches in Table 6. Nevertheless, the second category of approaches has certain limitations. First of all, the final 3D predictions are directly affected by the 2D estimator performances. Thus, if the 2D localization fails, the estimation of the 3D pose will also fail. In addition, these methods suffer from a lack of accuracy at low resolution due to the subsampling operations introduced in most 2D pose estimators. Second,

Table 5
Comparison of previous one-step single-person approaches on Human3.6 M based on Protocol 1. The used evaluation metric is MPJPE in mm.

3D one-step single-person approaches								
Model	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Li and Chan [77]	–	148.79	104.01	127.17	-	189.08	–	–
Tekin et al. [78]	–	129.06	91.43	121.68	–	162.17	–	–
Pavlakos et al. [79]	67.38	71.95	66.70	69.07	71.95	76.97	65.03	68.30
Sun et al. [35]	42.10	44.30	45.00	45.40	51.50	53.00	43.20	41.30
Yang et al. [81]	51.50	58.90	50.40	57.00	62.10	65.40	49.80	52.70
Sun et al. [80]	–	–	–	-	–	–	–	–
Rhodin et al. [82]	–	–	–	-	–	–	–	–
Model	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
Li and Chan [77]	–	–	–	–	146.59	77.60	–	–
Tekin et al. [78]	–	–	–	–	130.53	65.75	–	–
Pavlakos et al. [79]	83.66	96.51	71.74	65.83	74.89	59.11	63.24	71.90
Sun et al. [35]	59.30	73.30	51.00	44.00	48.00	38.30	44.80	48.30
Yang et al. [81]	69.20	85.20	57.40	58.40	43.60	60.10	47.70	58.60
Sun et al. [80]	–	–	–	–	–	–	–	40.60
Rhodin et al. [82]	–	–	–	–	–	–	–	131.70

Table 6
Comparison of previous two-step single-person approaches on Human3.6 M based on Protocol 1. The used evaluation metric is MPJPE in mm.

	3D two-step single-person approaches							
Model	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Mehta et al. [83]	62.60	78.10	63.40	72.50	88.30	93.80	63.10	74.80
Lin et al. [84]	58.02	68.16	63.25	65.77	75.26	93.05	61.16	65.71
Thom et al. [87]	64.98	73.47	76.82	86.43	86.28	110.67	68.93	74.79
Zhou et al. [85]	54.82	60.70	58.22	71.41	62.03	65.53	53.83	55.58
Martinez et al. [91]	37.70	44.40	40.30	42.10	48.20	54.90	44.40	42.10
Wang et al. [99]	49.20	55.50	53.60	53.40	63.80	67.70	50.20	51.90
Dabral et al. [86]	44.80	50.40	44.70	49.00	52.90	61.40	43.50	45.50
Pavllo et al. [94]	45.20	46.70	43.30	45.60	48.10	55.10	44.60	44.30
Kocabas et al. [88]	–	–	–	–	–	–	–	–
Gong et al. [89]	–	–	–	–	–	–	–	–
Iskakov et al. [96]	41.90	49.20	46.90	47.60	50.70	57.90	41.20	50.90
Liu et al. [100]	50.72	60.04	51.11	63.65	59.70	69.34	48.83	51.98
Xu et al. [101]	37.40	43.50	42.70	42.70	46.60	59.70	41.30	45.10
Ci et al. [92]	28.60	34.60	28.70	31.70	33.00	40.20	34.00	28.30
Remelli et al. [97]	27.30	32.10	25.00	26.50	29.30	35.40	28.80	31.60
Wandt et al. [98]	–	–	–	–	–	–	–	–

Model	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Walking
Mehta et al. [83]	106.60	138.70	78.80	73.90	82.00	55.80	59.60	80.5
Lin et al. [84]	98.65	127.68	70.37	68.17	72.94	50.63	57.74	73.10
Thom et al. [87]	110.19	173.91	84.95	85.78	86.26	71.36	73.14	88.39
Zhou et al. [85]	75.20	111.59	64.15	66.05	51.43	63.22	55.33	64.90
Martinez et al. [91]	54.60	58.00	45.10	46.40	47.60	36.40	40.4	45.50
Wang et al. [99]	70.30	87.50	57.70	51.50	58.60	44.60	47.20	57.80
Dabral et al. [86]	63.10	87.30	51.70	48.50	52.20	37.60	41.90	52.10
Pavllo et al. [94]	57.30	65.80	47.10	44.00	49.00	32.80	33.90	46.80
Kocabas et al. [88]	–	–	–	–	–	–	–	51.83
Gong et al. [89]	–	–	–	–	–	–	–	50.20
Iskakov et al. [96]	57.30	74.90	48.60	44.30	41.30	52.80	42.70	49.90
Liu et al. [100]	72.76	105.31	58.62	60.98	62.25	45.88	48.69	61.10
Xu et al. [101]	52.70	60.20	45.80	43.10	47.70	33.70	37.10	45.60
Ci et al. [92]	37.00	42.60	31.20	33.20	33.60	24.50	26.90	32.50
Remelli et al. [97]	36.40	31.70	31.20	29.90	26.90	33.70	30.40	30.20
Wandt et al. [98]	–	–	–	–	–	–	–	74.30

heatmaps are generally not accurate enough to identify two very close joints of the same type. Finally, the detection of two overlapped joints cannot be stored at the same location in 2D, resulting in erroneous estimation of the 3D pose. Moreover, the lifting of 2D joints to 3D from a single image remains a poorly posed problem.

For multi-person pipelines, top-down approaches achieved better results than bottom-up ones as they take advantage of the state-of-the-art person detectors and the single-person pose estimators. We summarize the results reached by these approaches on the Human3.6 M dataset in Table 7. However, similarly to the 2D top-down multi-person approaches, the computational complexity increases considerably with the number of persons in the image and may become too large in crowded scenes. Bottom-up approaches can solve these issues by ensuring a linear computational complexity but there is still a large gap in terms of performances between these two categories as illustrated in Table 8. Finally, most strategies attempt to generate more training data whether using multi-views, auto-encoders, or others. Also, the majority of these strategies are single-person.

Table 7
Comparison of previous top-down multi-person approaches on Human3.6 M.

Model	MPJPE	P-MPJPE
Top-down approaches		
Rogez et al. [102]	61.20	42.70
Moon et al. [103]	54.40	35.20
Dabral et al. [107]	65.20	–
Wang et al. [108]	48.60	30.5

6. Conclusion

The literature on human pose estimation has been widely developed since the introduction of deep learning techniques. In this paper, we presented a survey of recently published papers that address the human joint localization task from RGB images or video sequences. We suggested a taxonomy that organizes the mentioned approaches into various categories based on the general structure of each strategy. We focused on both 2D and 3D approaches. For each dimension space and based on the number of people in the input data, single-person and multi-person pipelines are presented.

We also provided a discussion which highlighted the advantages and the disadvantages of each category, as well as comparisons between different models and different pipelines. We also reviewed current datasets and evaluation metrics. Thus, this review can serve as a

Table 8
Comparison of previous multi-person approaches on MuPoTS-3D dataset. The used evaluation metric is 3DPCK.

Method	All people	Matched people
3D Top-down multi-person pose estimation		
Moon et al. [103]	81.80	82.50
Benzine et al. [105]	72.00	–
Dabral et al. [107]	71.30	74.20
Wang et al. [108]	82.00	–
3D Bottom-up multi-person pose estimation		
Mehta et al. [28]	65.00	69.80
Mehta et al. [110]	72.10	78.00
Kundu et al. [112]	74.00	75.80
Zhen et al. [113]	73.50	80.50

guideline for researchers interested in this field. It can also be used as a reference to study past models and develop new ones. As there are many challenges that need to be into consideration to create a robust pose estimation model, each paper proposed various techniques to address them. Consequently, combining some of these techniques can help improve the overall model performance.

Despite the variety of the proposed strategies, it can be seen that there is a trade-off between the inference run-time and the reached performances when using multi-person approaches. Despite the significant achievement realized in this field, there is still a need for further development to be used in real-world applications.

Currently, there is a vast selection of datasets. However, the unbalanced pose distribution can affect the model accuracy. To the best of our knowledge, no study proposed techniques to detect rare poses using this type of data. Besides, since one-stage 3D approaches can only be trained based on 3D annotations, the non-availability of a large-scale in-the-wild labeled dataset can lead to overfitting, which affects the pose estimator generalization. A possible improvement can be achieved by including techniques for either data augmentation, algorithmic data generation or simulation.

In conclusion, we are confident that this survey may help new researchers to better understand current pose estimation challenges and cover existing methods.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was enabled in part by support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2018-06233.

References

[1] M. Andriluka, S. Roth, B. Schiele, Pictorial structures revisited: People detection and articulated pose estimation, 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2009, pp. 1014–1021, https://doi.org/10.1109/cvprw.2009.5206754.

[2] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, 2011 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2011, pp. 1385–1392, https://doi.org/10.1109/CVPR.2011.59957411.

[3] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2014, pp. 1653–1660, https://doi.org/10.1109/CVPR.2014.214.

[4] N. Sarafianos, B. Boteanu, B. Ionescu, I.A. Kakadiaris, 3d human pose estimation: a review of the literature and analysis of covariates, Comput. Vis. Image Underst. 152 (2016) 1–20, https://doi.org/10.1016/j.cviu.2016.09.002.

[5] Y. Li, An Overview on 2d Multi-human Pose Estimation, 9, 2019.

[6] Q. Dang, J. Yin, B. Wang, W. Zheng, Deep learning based 2d human pose estimation: A survey, Tsinghua Science and Technology 24 (2019) 663–676, https://doi.org/10.26599/TST.2018.9010100.

[7] T.L. Munea, Y.Z. Jembre, H.T. Weldegebriel, L. Chen, C. Huang, C. Yang, The progress of human pose estimation: a survey and taxonomy of models applied in 2d human pose estimation, IEEE Access 8 (2020) 133330–133348, https://doi.org/10.1109/ACCESS.2020.3010248.

[8] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, M. Shah, Deep learning-based human pose estimation: A survey, arXiv preprint (2020) https://doi.org/10.26599/TST.2018.9010100.

[9] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, 2014 IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 3686–3693, https://doi.org/10.1109/CVPR.2014.471.

[10] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, et al., Ai challenger: A large-scale dataset for going deeper in image understanding, arXiv preprint abs/1711.06475 (2017) 1–11.

[11] S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, British Machine Vision Conference 2010, pp. 12.1–12.11, https://doi.org/10.5244/C.24.12.

[12] S. Johnson, M. Everingham, Learning effective human pose estimation from inaccurate annotation, 2011 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2011, pp. 1465–1472, https://doi.org/10.1109/CVPR.2011.5995318.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll'ar, C.L. Zitnick, Microsoft coco: Common objects in context, European Conference on Computer Vision, Springer 2014, pp. 740–755, https://doi.org/10.1007/978-3-319-10602-1_48.

[14] B. Sapp, B. Taskar, Modec: Multimodal decomposable models for human pose estimation, 2013 IEEE Conference on Computer Vision and Pattern Recognition 2013, pp. 3674–3681, https://doi.org/10.1109/CVPR.2013.471.

[15] X. Liang, K. Gong, X. Shen, L. Lin, Look into person: Joint body parsing & pose estimation network and a new benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41, 2018, pp. 871–885, https://doi.org/10.1109/TPAMI.2018.2820063.

[16] F. Xia, P. Wang, X. Chen, A.L. Yuille, Joint multi-person pose estimation and semantic part segmentation, 2017 IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 6769–6778, https://doi.org/10.1109/CVPR.2017.644.

[17] M. Everingham, L. Van Gool, C. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results, http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html 2011.

[18] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, C. Lu, Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 10855–10864, https://doi.org/10.1109/CVPR.2019.01112.

[19] W. Zhang, M. Zhu, K.G. Derpanis, From actemes to action: A stronglysupervised representation for detailed action understanding, 2013 IEEE International Conference on Computer Vision 2013, pp. 2248–2255, https://doi.org/10.1109/ICCV.2013.280.

[20] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1325–1339, https://doi.org/10.1109/TPAMI.2013.248.

[21] L. Sigal, A.O. Balan, M.J. Black, Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Int. J. Comput. Vis. 87 (2009) 4–27, https://doi.org/10.1007/s11263-009-0273-6.

[22] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3d human pose estimation in the wild using improved cnn supervision, 2017 International Conference on 3D Vision (3DV), IEEE 2017, pp. 506–516, https://doi.org/10.1109/3DV.2017.00064.

[23] S. Antol, C.L. Zitnick, D. Parikh, Zero-shot learning via visual abstraction, European Conference on Computer Vision, Springer 2014, pp. 401–416, https://doi.org/10.1007/978-3-319-10593-2_27.

[24] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, B. Schiele, Posetrack: A benchmark for human pose estimation and tracking, 2018 IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 5167–5176, https://doi.org/10.1109/CVPR.2018.00542.

[25] J. Charles, T. Pfister, M. Everingham, A. Zisserman, Automatic and efficient human pose estimation for sign language videos, Int. J. Comput. Vis. 110 (2013) 70–90, https://doi.org/10.1007/s11263-013-0672-6.

[26] T. Pfister, K. Simonyan, J. Charles, A. Zisserman, Deep convolutional neural networks for efficient pose estimation in gesture videos, Asian Conference on Computer Vision, Springer 2014, pp. 538–552, https://doi.org/10.1007/978-3-319-16865-4_35.

[27] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M.J. Black, Towards understanding action recognition, 2013 IEEE International Conference on Computer Vision 2013, pp. 3192–3199, https://doi.org/10.1109/ICCV.2013.396.

[28] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, C. Theobalt, Single-shot multi-person 3d pose estimation from monocular rgb, 2018 International Conference on 3D Vision (3DV), IEEE 2018, pp. 120–130, https://doi.org/10.1109/3DV.2018.00024.

[29] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, et al., Panoptic studio: a massively multiview system for social interaction capture, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2017) 190–204, https://doi.org/10.1109/TPAMI.2017. 2782743.

[30] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, S. Ilic, 3d pictorial structures for multiple human pose estimation, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, 1669–1676https://doi.org/10.1109/CVPR.2014.216.

[31] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, R. Cucchiara, Learning to detect and track visible and occluded body joints in a virtual world, European Conference on Computer Vision (ECCV) 2018, pp. 430–446, https://doi.org/10.1007/978-3-030-01225-0_27.

[32] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, G. Pons-Moll, Recovering accurate 3d human pose in the wild using imus and a moving camera, European Conference on Computer Vision (ECCV) 2018, pp. 601–617, https://doi.org/10.1007/978-3-030-01249-6_37.

[33] M. Eichner, M. Marin-Jimenez, A. Zisserman, V. Ferrari, Articulated human pose estimation and search in (almost) unconstrained still images, ETH Zurich, D-ITET, BIWI, Technical Report No 272, 2010.

[34] J. Carreira, P. Agrawal, K. Fragkiadaki, J. Malik, Human pose estimation with iterative error feedback, 2016 IEEE Conference on Computer 1310 Vision and Pattern Recognition 2016, pp. 4733–4742, https://doi.org/10.1109/CVPR.2016.512.

[35] X. Sun, J. Shang, S. Liang, Y. Wei, Compositional human pose regression, 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 2602–2611, https://doi.org/10.1109/ICCV.2017.284.

[36] D.C. Luvizon, H. Tabia, D. Picard, Human pose regression by combining indirect part detection and contextual information, Comput. Graph. 85 (2019) 15–22, https://doi.org/10.1016/j.cag.2019.09.002.

[37] F. Zhang, X. Zhu, H. Dai, M. Ye, C. Zhu, Distribution-aware coordinate representation for human pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020, pp. 7093–7102, https://doi.org/10.1109/cvpr42600.2020.00712.

[38] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, European Conference on Computer Vision, Springer 2016, pp. 483–499, https://doi.org/10.1007/978-3-319-46484-8_29.

[39] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, J. Wang, Human pose estimation using global and local normalization, 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 5599–5607, https://doi.org/10.1109/ICCV.2017.597.

[40] L. Ke, M.-C. Chang, H. Qi, S. Lyu, Multi-scale structure-aware networ for human pose estimation, European Conference on Computer Vision 2018, pp. 713–728, https://doi.org/10.1007/978-3-030-01216-8_44.

[41] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 1281–1290, https://doi.org/10.1109/ICCV.2017.144.

[42] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (2020) 139–144, https://doi.org/10.1007/978-1-4842-3679-6_8.

[43] C.-J. Chou, J.-T. Chien, H.-T. Chen, Self adversarial training for human pose estimation, 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE 2018, pp. 17–30, https://doi.org/10.23919/APSIPA.2018.8659538.

[44] Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial posenet: A structure-aware convolutional network for human pose estimation, 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 1212–1221, https://doi.org/10.1109/ICCV.2017.137.

[45] J. Wang, S. Jin, W. Liu, W. Liu, C. Qian, P. Luo, When human pose estimation meets robustness: Adversarial algorithms and benchmarks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021, pp. 11855–11864.

[46] X. Nie, J. Feng, Y. Zuo, S. Yan, Human pose estimation with parsing induced learner, 2018 IEEE Conference on Computer Vision and Patter Recognition 2018, pp. 2100–2108, https://doi.org/10.1109/CVPR.2018.00224.

[47] Z. Su, M. Ye, G. Zhang, L. Dai, J. Sheng, Cascade Feature Aggregation for Human Pose Estimation, arXiv: Computer Vision and Pattern Recognition, 2019.

[48] A. Bulat, J. Kossaifi, G. Tzimiropoulos, M. Pantic, Toward fast and accurate human pose estimation via soft-gated skip connections, 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) 2020, pp. 8–15, https://doi.org/10.1109/FG47880.2020.00014.

[49] B. Artacho, A. Savakis, Unipose: Unified human pose estimation in single images and videos, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 7035–7044, https://doi.org/10.1109/cvpr42600. 2020.00706.

[50] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2015) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.

[51] K. He, G. Gkioxari, P. Doll'ar, R. Girshick, Mask r-cnn, 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 2961–2969, https://doi.org/10.1109/TPAMI.2018.2844175.

[52] T.-Y. Lin, P. Doll'ar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, pp. 2117–2125, https://doi.org/10.1109/CVPR.2017.106.

[53] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, 2016 IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 4724–4732, https://doi.org/10.1109/CVPR.2016.511.

[54] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, Towards accurate multi-person pose estimation in the wild, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, pp. 4903–4911, https://doi.org/10.1109/CVPR.2017.395.

[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 770–778, https://doi.org/10.1109/cvpr.2016.90.

[56] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, Rmpe: Regional multi-person pose estimation, 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 2334–2343, https://doi.org/10.1109/ICCV.2017.256.

[57] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, pp. 7103–7112, https://doi.org/10.1109/CVPR.2018.00742.

[58] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, European Conference on Computer Vision (ECCV) 2018, pp. 466–481, https://doi.org/10.1007/978-3-030-01231-1_29.

[59] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019, pp. 5693–5703, https://doi.org/10.1109/CVPR.2019.00584.

[60] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, J. Sun, Rethinking on multi-stage networks for human pose estimation, arXiv: Computer Vision and Pattern Recognition abs/1901.00148 (2019) 1–10.

[61] J. Huang, Z. Zhu, F. Guo, G. Huang, The devil is in the details: Delving into unbiased data processing for human pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020, pp. 5699–5708, https://doi.org/10.1109/cvpr42600.2020.00574.

[62] Y. Bin, Z.-M. Chen, X.-S. Wei, X. Chen, C. Gao, N. Sang, Structure-aware human pose estimation with graph convolutional networks, Pattern Recogn. 106 (2020) 107410, https://doi.org/10.1016/j.patcog.2020.107410.

[63] F. Zhong, M. Li, K. Zhang, J. Hu, L. Liu, Dspnet: a low computationalcost network for human pose estimation, Neurocomputing 423 (2021) 327–335, https://doi.org/10.1016/j.neucom.2020.11.003.

[64] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P.V. Gehler, B. Schiele, Deepcut: Joint subset partition and labeling for multi person pose estimation, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 4929–4937, https://doi.org/10.1109/CVPR.2016.533.

[65] R. Girshick, Fast r-cnn, 2015 IEEE International Conference on Computer Vision (ICCV) 2015, pp. 1440–1448, https://doi.org/10.1109/ICCV.2015169.

[66] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, Deepercut: A deeper, stronger, and faster multi-person pose estimation model, European Conference on Computer Vision, Springer 2016, pp. 34–50, https://doi.org/10.1007/978-3-319-46466-4_3.

[67] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, 2017 IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 7291–7299, https://doi.org/10.1109/CVPR.2017.143.

[68] D. Osokin, Real-time 2d multi-person pose estimation on cpu: Lightweight openpose, The 8th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM, 2019 https://doi.org/10.5220/0007555407440748.

[69] K. Simonyan, A. Zisserman, Very deep convolutional networks for largescale image recognition, arXiv preprint abs/1409.1556 (2014) 1–14.

[70] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint abs/1704.04861 (2017) 1–9.

[71] S. Kreiss, L. Bertoni, A. Alahi, Pifpaf: Composite fields for human pose estimation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019, pp. 11977–11986, https://doi.org/10.1109/CVPR.2019.01225.

[72] B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang, L. Zhang, Bottomup higherresolution networks for multi-person pose estimation, arXiv preprint abs/1908.10357 (2019) 1–10.

[73] H.-P. Cheng, F. Liang, M. Li, B. Cheng, F. Yan, H. Li, V. Chandra, Y. Chen, Scalenas: One-shot learning of scale-aware representations for visual recognition, arXiv preprint abs/2011.14584 (2020) 1–12.

[74] Z. Luo, Z. Wang, Y. Huang, T. Tan, E. Zhou, Rethinking the heatmap regression for bottom-up human pose estimation, arXiv preprint abs/2012.15175 (2020) 1–10.

[75] B. Artacho, A. Savakis, Omnipose: A multi-scale framework for multi-person pose estimation, arXiv preprint abs/2103.10180 (2021) 1–10.

[76] Z. Geng, K. Sun, B. Xiao, Z. Zhang, J. Wang, Bottom-up human pose estimation via disentangled keypoint regression, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 14676–14686.

[77] S. Li, A.B. Chan, 3d human pose estimation from monocular images with deep convolutional neural network, Asian Conference on Computer Vision, Springer 2014, pp. 332–347, https://doi.org/10.1007/978-3-319-16808-1_23.

[78] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, P. Fua, et al., arXiv preprint (2016) https://doi.org/10.5244/c.30.130.

[79] G. Pavlakos, X. Zhou, K.G. Derpanis, K. Daniilidis, Coarse-to-fine volumetric prediction for single-image 3d human pose, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, pp. 7025–7034, https://doi.org/10.1109/CVPR.2017.139.

[80] X. Sun, B. Xiao, F. Wei, S. Liang, Y. Wei, Integral human pose regression, European Conference on Computer Vision (ECCV) 2018, pp. 529–545, https://doi.org/10.1007/978-3-030-01231-1_33.

[81] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, X. Wang, 3d human pose estimation in the wild by adversarial learning, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, pp. 5255–5264, https://doi.org/10.1109/CVPR.2018.00551.

[82] H. Rhodin, M. Salzmann, P. Fua, Unsupervised geometry-aware representation for 3d human pose estimation, European Conference on Computer Vision (ECCV) 2018, pp. 750–767, https://doi.org/10.1007/978-3-030-01249-6_46.

[83] S. Mehta, O. Sridhar, H. Sotnychenko, M. Rhodin, H.-P. Shafiei, W. Seidel, D. Xu, C. Casas, Theobalt, Vnect: real-time 3d human pose estimation with a single rgb camera, ACM Transactions on Graphics (TOG) 36 (2017) 1–14, https://doi.org/10.1145/3072959.3073596.

[84] M. Lin, L. Lin, X. Liang, K. Wang, H. Cheng, Recurrent 3d pose sequence machines, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, pp. 810–819, https://doi.org/10.1109/CVPR.2017.588.

[85] X. Zhou, Q. Huang, X. Sun, X. Xue, Y. Wei, Towards 3d human pose estimation in the wild: a weakly-supervised approach, 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 398–407, https://doi.org/10.1109/ICCV.2017.51.

[86] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, A. Jain, Learning 3d human pose from structure and motion, European Conference on Computer Vision (ECCV) 2018, pp. 668–683, https://doi.org/10.1007/978-3-030-01240-3_41.

[87] D. Tome, C. Russell, L. Agapito, Lifting from the deep: Convolutional 3d pose estimation from a single image, 2017 IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 2500–2509, https://doi.org/10.1109/CVPR.2017.603.

[88] M. Kocabas, S. Karagoz, E. Akbas, Self-supervised learning of 3d human pose using multi-view geometry, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 1077–1086, https://doi.org/10.1109/CVPR.2019.00117.

[89] K. Gong, J. Zhang, J. Feng, Poseaug: A differentiable pose augmentation framework for 3d human pose estimation, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 8575–8584.

[90] B. Wandt, B. Rosenhahn, Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 7782–7791.
[91] J. Martinez, R. Hossain, J. Romero, J.J. Little, A simple yet effective baseline for 3d human pose estimation, 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 2640–2649, https://doi.org/10.1109/ICCV.2017.288.
[92] H. Ci, X. Ma, C. Wang, Y. Wang, Locally connected network for monocular 3d human pose estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1, https://doi.org/10.1109/TPAMI.2020.3019139.
[93] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, Advances in Neural Information Processing Systems, vol. 29, Curran Associates, Inc 2016, pp. 3844–3852.
[94] D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 3d human pose estimation in video with temporal convolutions and semi-supervised training, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019, pp. 7753–7762, https://doi.org/10.1109/CVPR.2019.00794.
[95] K. Takahashi, D. Mikami, M. Isogawa, H. Kimata, Human pose as calibration pattern; 3d human pose estimation with multiple unsynchronized and uncalibrated cameras, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2018, pp. 1775–1782, https://doi.org/10.1109/CVPRW.2018.00230.
[96] K. Iskakov, E. Burkov, V. Lempitsky, Y. Malkov, Learnable triangulation of human pose, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) 2019, pp. 7718–7727, https://doi.org/10.1109/ICCV.2019.00781.
[97] E. Remelli, S. Han, S. Honari, P. Fua, R. Wang, Lightweight multiview 3d pose estimation through camera-disentangled representation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020, pp. 6040–6049, https://doi.org/10.1109/CVPR42600.2020.00608.
[98] B. Wandt, M. Rudolph, P. Zell, H. Rhodin, B. Rosenhahn, Canonpose: Self-supervised monocular 3d human pose estimation in the wild, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 13294–13304.
[99] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, L. Ma, Drpose3d: Depth ranking in 3d human pose estimation, arXiv preprint (2018) https://doi.org/10.24963/ijcai.2018/136.
[100] J. Liu, H. Ding, A. Shahroudy, L.-Y. Duan, X. Jiang, G. Wang, A.C. Kot, Feature boosting network for 3d pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2019) 494–501, https://doi.org/10.1109/TPAMI.2019.2894422.
[101] J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, W. Zhang, Deep kinematics analysis for monocular 3d human pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020, pp. 899–908, https://doi.org/10.1109/cvpr42600.2020.00098.
[102] G. Rogez, P. Weinzaepfel, C. Schmid, Lcr-net++: multi-person 2d and 3d pose detection in natural images, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2019) 1146–1161, https://doi.org/10.1109/TPAMI.2019. 2892985.
[103] G. Moon, J.Y. Chang, K.M. Lee, Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image, 2019 IEEE/CVF International Conference on Computer Vision 2019, pp. 10133–10142, https://doi.org/10.1109/ICCV.2019.01023.
[104] J. Dong, W. Jiang, Q. Huang, H. Bao, X. Zhou, Fast and robust multiperson 3d pose estimation from multiple views, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 7792–7801, https://doi.org/10.1109/CVPR.2019.00798.
[105] A. Benzine, F. Chabot, B. Luvison, Q.C. Pham, C. Achard, Pandanet: Anchor-based single-shot multi-person 3d pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 6856–6865, https://doi.org/10.1109/cvpr42600.2020.00689.
[106] H. Tu, C. Wang, W. Zeng, Voxelpose: towards multi-camera 3d human pose estimation in wild environment, European Conference on Computer Vision (ECCV), 2020https://doi.org/10.1007/978-3-030-58452-8_12.
[107] R. Dabral, N.B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, A. Jain, Multi-person 3d human pose estimation from monocular images, 2019 International Conference on 3D Vision (3DV), IEEE 2019, pp. 405–414, https://doi.org/10.1109/3DV.2019.00052.
[108] C. Wang, J. Li, W. Liu, C. Qian, C. Lu, Hmor: Hierarchical multiperson ordinal relations for monocular multi-person 3d pose estimation, European Conference on Computer Vision, Springer 2020, pp. 242–259, https://doi.org/10.1007/978-3-030-58580-8_15.
[109] H. Chen, P. Guo, P. Li, G.H. Lee, G. Chirikjian, Multi-person 3d pose estimation in crowded scenes based on multi-view geometry, European Conference on Computer Vision, Springer 2020, pp. 541–557, https://doi.org/10.1007/978-3-030-58580-8_32.
[110] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, C. Theobalt, Xnect: real-time multiperson 3d motion capture with a single rgb camera, ACM Transactions on Graphics (TOG) 39 (2020) https://doi.org/10.1145/3386569.3392410 82–1.
[111] M. Fabbri, F. Lanzi, S. Calderara, S. Alletto, R. Cucchiara, Compressed volumetric heatmaps for multi-person 3d pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 7204–7213, https://doi.org/10.1109/CVPR42600.2020.00723.
[112] J.N. Kundu, A. Revanur, G.V. Waghmare, R.M. Venkatesh, R.V. Babu, Unsupervised cross-modal alignment for multi-person 3d pose estimation, European Conference on Computer Vision, Springer 2020, pp. 35–52, https://doi.org/10.1007/978-3-030-58601-0_3.
[113] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao, X. Zhou, Smap: Single-shot multi-person absolute 3d pose estimation, European Conference on Computer Vision, Springer 2020, pp. 550–566, https://doi.org/10.1007/978-3-030-58555-6_33.