TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 03/10 pp663–676
DOI: 10.26599/TST.2018.9010100
Volume 24, Number 6, December 2019

Deep Learning Based 2D Human Pose Estimation: A Survey

Qi Dang, Jianqin Yin*, Bin Wang, and Wenqing Zheng

Abstract: Human pose estimation has received significant attention recently due to its various applications in the real world. As the performance of the state-of-the-art human pose estimation methods can be improved by deep learning, this paper presents a comprehensive survey of deep learning based human pose estimation methods and analyzes the methodologies employed. We summarize and discuss recent works with a methodology-based taxonomy. Single-person and multi-person pipelines are first reviewed separately. Then, the deep learning techniques applied in these pipelines are compared and analyzed. The datasets and metrics used in this task are also discussed and compared. The aim of this survey is to make every step in the estimation pipelines interpretable and to provide readers a readily comprehensible explanation. Moreover, the unsolved problems and challenges for future research are discussed.

Key words: human pose estimation; deep learning; computer vision

1 Introduction

Human pose estimation is widely used in human-computer interactions, gaming, virtual reality, video surveillance, sports analysis, and medical assistance, making it a highly popular research topic in the field of computer vision[1–4]. Human pose estimation aims to automatically locate the human body parts from images or videos. In a simple case, as shown in Fig. 1a, where only one person is in the image or the position of the person is given, a single-person algorithm should be performed to locate the human parts, such as the top of the head, the center of the neck, the left/right elbows, and the left/right shoulders[6–8]. In more general cases, as shown in Fig. 1b, where the number and the position of persons in an image are unknown, the multi-person pose estimation algorithms are performed[2,3,9]. All the human parts in an image should be detected and the keypoints of the same person, even in a crowded scene, should be associated.

To solve the problem of human pose estimation, a number of approaches have been proposed in the literature. Hand-crafted features have been used in early works. These hand-crafted features, such as HOG (Histogram of Oriented Gradient)[10–14] (Fig. 2) and Edgelet[15], are insufficient in determining the accurate locations of body parts. By contrast, deep learning based methods are capable of extracting more sufficient features from meta data. Such methods have yielded excellent results and outperformed non-deep state-of-the-art methods[10–15] with a big margin. Although the utilization of deep learning in pose estimation field is relatively new, numerous outstanding works on this topic have been conducted. However, to the best of our knowledge, no previous survey has reviewed existing works on deep learning for human pose estimation. Hence, the objective of the current paper is to provide a comprehensive overview of state-of-the-art deep learning based two-dimensional (2D) human pose estimation methodologies and provide further research

• Qi Dang and Jianqin Yin are with Automation School, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: dangqi213@163.com; jqyin@bupt.edu.cn.
• Qi Dang and Bin Wang are with State Key Lab. of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China. E-mail: wangbinth@tsinghua.edu.cn.
• Wenqing Zheng is with School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: zhengwenqing@bupt.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2018-03-05; revised: 2018-04-30; accepted: 2018-05-03

(a)

(b)

Fig. 1 Examples of pose estimation results. (a) Single person pose estimation results from Ref. [5]. (b) Multi-person pose estimation results from Ref. [3].

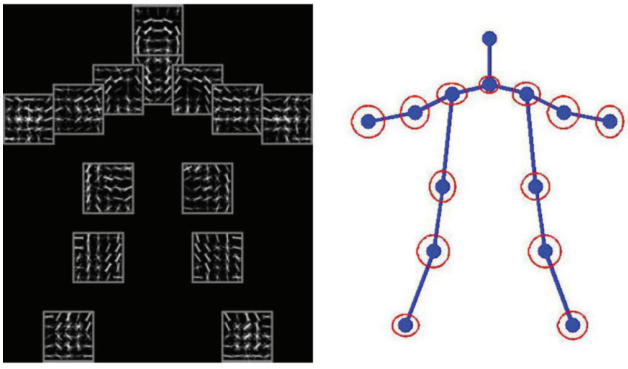


Fig. 2 Example of HOG features for keypoints detection[10].

trends for readers. We hope that readers can gain inspiration from our paper.

The tree-structured taxonomy that our survey follows is illustrated in Fig. 3. For the current paper, we have chosen methodology-based taxonomy. All human pose estimation methodologies are first classified into two categories: single-person pose estimation approaches and multi-person pose estimation approaches. On the one hand, the single-person approaches detect the pose of a certain person in an image. Due to the given position information, the objective of single-person approaches is to find the keypoint position in that area, so it is essential to solve a regression problem, in which

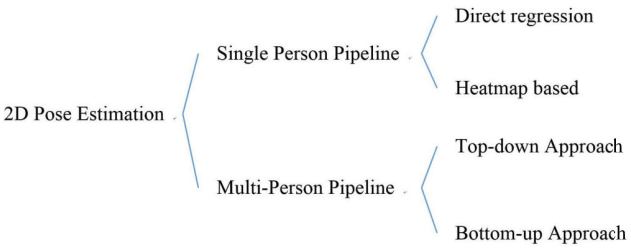


Fig. 3 Taxonomy of this review.

the amount of keypoints is implicitly given. On the other hand, the objective of multi-person approaches is to solve an unconstrained problem, because the number and position of persons are unknown.

A single-person pipeline is classified into two types depending on the way they predict keypoints: direct regression-based approaches and heatmap-based approaches. The former utilize the output feature maps to regress keypoints directly, whereas the latter generate heatmaps (the pixel value in heatmap indicates the keypoint existence probability in that position) first and predict keypoints based on the heatmaps. Multi-person approaches can be classified into two categories according to their methodology: top-down approaches and bottom-up approaches. Top-down approaches are roughly divided into two steps: human detection and single-person keypoint estimation. Bottom-up methods have similar steps with a reversed order: the first step is to locate all the keypoints in an image and the second step is to group these keypoints according to the person they belong to.

There have been other surveys related to pose estimation. For example, Guo et al.[16] reviewed the deep learning methods and took pose estimation as part of it, but they focused on the applications of deep learning for computer vision and just summarized the pose estimation methods briefly. Poppe[17] reviewed early methods for vision-based human motion analysis. Liu et al.[18] investigated body parts parsing based methods for human pose estimation, but most of the methods they reviewed are based on hand-crafted features. Zhang et al.[19] and Gong et al.[20] also surveyed human pose estimation methods, but their surveys did not focus on deep learning based approaches. The pose estimation methods for particular human part, such as hand and head, have been reviewed in Refs. [21, 22]. Asadi-Aghbolaghi et al.[23] surveyed deep learning based approaches for action and gesture recognition in image sequences, and discussed deep learning techniques applied to action and gesture

recognition.

However, most of the previous surveys have focused on introducing or summarizing the traditional methodologies of human pose estimation and did not fully discuss nor analyze the deep learning based methods. In the current paper, the comparisons among different deep learning based methods are presented. The structures and tricks applied in these methods are discussed and analyzed. The goal of this paper is to analyze the key procedures of various proposed approaches and make readers understand how deep learning can be exploited in human pose estimation tasks.

2 Single-Person Pipeline

Single-person approaches estimate human pose in an image or a video under the condition that the position of the human is given. Generally, the position and rough scale of a person or the bounding box of a person are provided before estimation. Early works model human parts as a stickman (Fig. 4), but recent works model it as key joints because such joints are connected naturally and have more accurate positions. The objective of deep learning based single-person approaches is to locate keypoints of human parts. There are two typical frameworks for single-person pipeline: (1) The first one directly regresses keypoints from the features, and we call it direct regression based framework (Fig. 5a). (2) The other generates heatmaps first and inferences keypoint location via heatmaps. We call this heatmap-based framework (Fig. 5b).

2.1 Direct regression based framework

Several works are based on direct regression framework. Toshev and Szegedy[5] proposed a cascaded DNN regressor to predict human keypoints directly. However, it is difficult to learn mapping directly from feature maps without other procedures. Carreira et al.[25] used a self-correcting model. By feeding back error predictions, the predicted keypoint locations are refined progressively. Sun et al.[26] proposed a structure-aware approach called "compositional pose regression". Unlike other related works, this approach re-parameterizes pose representation using bones instead of joints, which is more primitive, stable, and easier to learn. Long-range interactions between bones are encoded by a compositional loss function. Luvizon et al.[27] proposed Soft-argmax to convert heatmaps to coordinates in a fully differentiable fashion. A keypoint error distance based loss function and a context-based structure are utilized in their end-to-end trainable

Fig. 4 Example of stickman annotations[24].

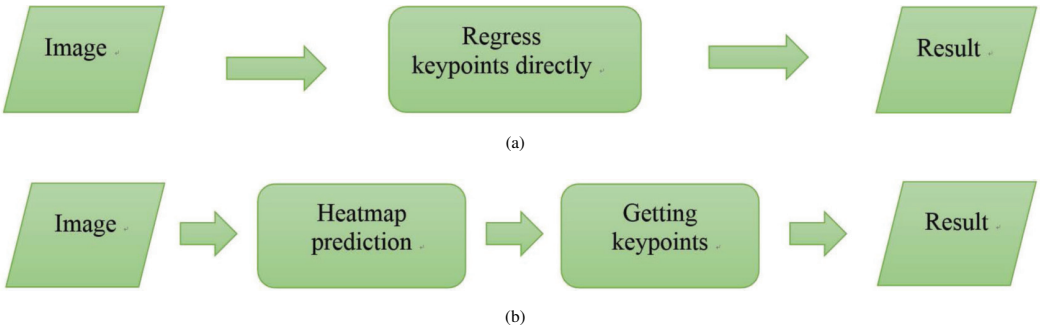


Fig. 5 Framework of the pipeline for single-person pose estimation. (a) Heatmap-based framework: There are two steps, generating heatmaps and regressing keypoints. (b) One step framework: There is just one step, human keypoints are regressed directly.

network, enabling it to achieve comparable results to the state-of-the-art heatmap-based framework.

2.2 Heatmap-based framework

As shown in Fig. 6, numerous works have employed the heatmap-based framework. Some works exploited human priors in their models. For example, Chen and Yuille[6] used the graphical model with pairwise relations learned by DCNN (parts type and pairwise parts relationships). In another study, Chen et al.[8] incorporated the priors of human bodies by employing the training strategy of conditional Generative Adversarial Networks (GANs)[28].

Network structure design has always been a theme of deep learning based approaches. Convolutional Pose Machines (CPM)[4] regress the heatmaps in multiple stages and use intermediate supervision to avoid the vanishing gradient. Newell et al.[29] designed a novel network structure called "stacked hourglass". Repeated bottom-up, top-down processing with intermediate supervision is proven to be critical for improving human pose detection performance. Chu et al.[30] built their baseline model based on stacked hourglass. They employed a multi-context attention mechanism to make the model more robust and more accurate. They also modified the structure of stacked hourglass by coupling the hourglass residual unit in it.

The relationship between 2D and three-dimensional (3D) keypoint detection was explored as well. Martinez et al.[7] proposed using 2D keypoints directly to predict 3D keypoints with deep neural networks. Their experiment results revealed that the 2D detection is one of the main causes of errors in 3D human pose estimation.

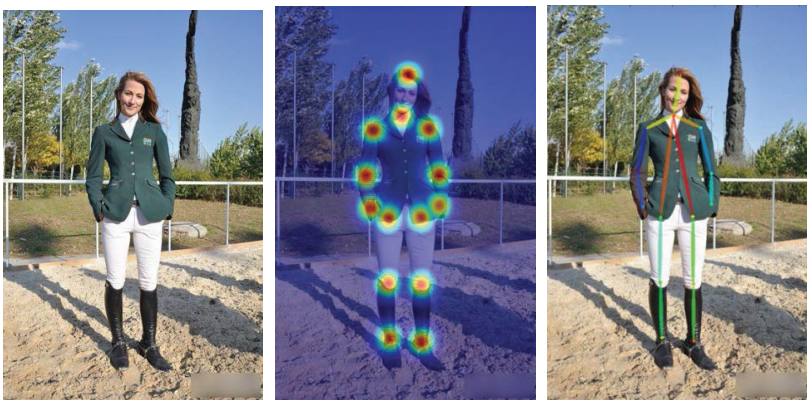
2.3 Discussion

2.3.1 Which framework is better, direct regression based framework or heatmap-based framework?

Earlier works[5,31] and a few recent works[25–27] have attempted to regress the coordinates of keypoints directly. The direct regression of joint positions is highly non-linear and has difficulty in learning mapping[32,33]. Furthermore, it cannot be applied to a multi-person case (bottom-up approaches or a single-detection box containing more than one person). By contrast, the heatmap-based framework regresses heatmaps first. Heatmaps can be visualized to enhance human understanding and model more complicated cases. However, if these particular techniques are combined[25–27], direct regression can be more reliable and has some merits. When direct regression is applied, the final result can be obtained in an end-to-end fashion without handling heatmaps. Moreover, it can be applied to 3D scenarios without too many changes. Additionally, the precision of predicting results relies on heatmap resolution, which requires a high memory consumption[27]. Therefore, no absolute conclusion for this question can be found, and each framework has its advantages and disadvantages. The comparison is shown in Table 1.

3 Multi-Person Pipeline

Compared with the single-person pipeline, the multi-person pipeline is more difficult because neither the number nor the position of the person is given. Keypoint detection and human location are two core problems in this task. To solve these two problems, two



(a) (b) (c)

Fig. 6 An example of heatmap-based single-person pipeline with heatmap. (a) Original image, (b) heatmap generated by estimator, and (c) detection result.

Table 1 Comparison between direct regression based framework and heatmap-based framework.

Framework	Advantage	Disadvantage
Direct regression based	Quick and direct, trained with an end-to-end fashion. Can be applied to 3D scenarios without much changes.	Difficult to learn mapping. Cannot be applied to multi-person case.
Heatmap-based	Easy to be visualized. Can be applied to complicated case.	High memory consumption for getting high resolution heat map. Hard to be extended to 3D scenarios.

popular pipelines have been proposed: (1) top-down pipeline and (2) bottom-up pipeline.

3.1 Top-down pipeline

The framework of the top-down pipeline is shown in Fig. 7. The first step of the top-down approach is to detect all persons from a given image, after which single-person approaches are performed in each detected bounding box. An illustration of the top-down pipeline is shown in Fig. 8. Aside from person detection, top-down approaches have more process, which may employ context information from the whole image.

Toshev and Szegedy[5] proposed the first deep learning based top-down method using the FLIC dataset[34]. Human pose estimation is regarded as a key point regression problem in their work. The face-based body detector is used to first estimate the rough position of the person, then a multi-stage cascade DNN based joint coordinate regressor is employed to regress the joint coordinates directly. Data augmentation was explored in a previous work[35]. Radosavovic et al.[35] exploited omni-supervised learning, which employed all the available data, labeled and unlabeled, to train the model. This work proved that the state-of-the-art human pose detectors are accurate enough to apply self-training techniques to challenging real-world data.

Human detection and human box alignment are studied in Refs. [36, 37]. Fang et al.[36] noticed that single-person pose estimation is sensitive to human detection. To solve this problem, they employed Symmetric Spatial Transformer Network (SSTN) with parallel Single-Person Pose Estimator (SPPE) to extract a high-quality single-person region. Mask R-CNN[37] simultaneously predicts human bounding box and human keypoints, which makes the detection faster by sharing the features of ConvNet. Moreover, RoI alignment enables a more accurate feature map cropping method. No human skeleton priors are combined in this work. Adding such skeleton priors

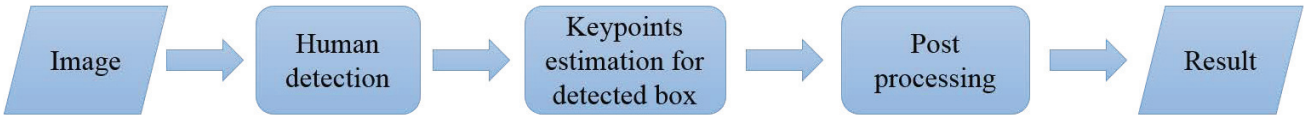


Fig. 7 Framework of top-down pipeline.

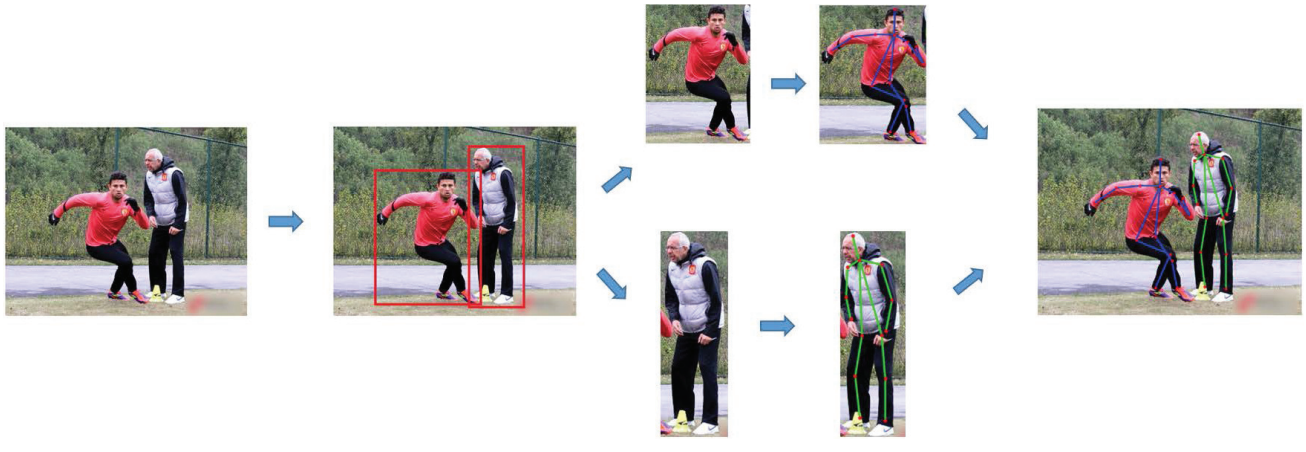


Fig. 8 An illustration of top down pipeline. (a) Input image, (b) two persons detected by human detector, (c) cropped single person image, (d) single person pose detection result, and (e) multi-person pose detection result.

may further boost the accuracy of Mask R-CNN.

Some works[1,38,39] focused on keypoint estimation within the human detection box. For example, Iqbal and Gall[38] considered multi-person pose estimation as a joint-to-person association problem. The aim of their method is to solve the occlusion problem. The redundant person parts in the same detection box are eliminated by associating the detected keypoints locally. This work is based on Ref. [9], but performs faster by importing locally associated mechanisms. Papandreou et al.[1] proposed a method that predicts joint dense heatmaps and position offsets simultaneously, after which these two outputs are aggregated to obtain highly localized keypoint positions. Chen et al.[39] proposed a network structure dubbed Cascaded Pyramid Network, which consists of two parts: GlobalNet and RefineNet. The former can catch a good feature representation, whereas the latter is employed to address the "hard" examples.

Post-processing methods have been proposed in Refs. [1,36]. A data-driven pose Non-Maximum Suppression (NMS) is proposed in Ref. [36] to solve the occlusion problem and a pose-guided proposal generator is used to perform data augmentation. However, serious occlusion or misdetection is still a big challenge for this approach. Pose rescoring and pose-based NMS method were used in Ref. [1] to eliminate the false positives while keeping the true positives.

3.2 Discussion for top-down approaches

3.2.1 Does the human detector matter in top-down human pose estimation?

The first step of top-down approach is human detection. The most popular human detectors used in human pose estimation are based on the Faster R-CNN structure because it is an off-the-shelf detector with high performance. Faster R-CNN has many variants with different base networks (VGG[40], ResNet[41], and Inception- ResNet[42]) and different extended structures (FPN[43]). These variants have different levels of accuracy, inference time, and computing complexity. Generally, the more accurate the detection result is, the more complex the network is assumed to be. Therefore, a trade-off among accuracy, memory, and time should be considered.

Some works have compared the performance of the human pose estimator with different human detectors[1,39]. The results of most works show that the accuracy of the human pose estimator increases with a better human detector. Meanwhile, as seen in Fig. 9, the result of Ref. [39] shows that the human pose estimator achieves a big gain from a better human detector when detector is of low performance. With the increase of the human detector's Average Precision (AP), the AP of the human pose estimator increases slower. When the human detectors achieve very high accuracy (ensemble of many high-performance human detectors), the accuracy of their pose estimation network could no longer increase. The possible reason is that the persons who are not detected by the human detectors are also difficult examples for the human pose estimator.

In other words, the human detector matters when its performance is ordinary, but it does not matter when it has already achieved high performance. The gain of the pose estimator is very little with higher human detection AP, especially when a human detector is already accurate enough.

3.2.2 NMS

NMS is a common method to suppress redundant detections. This technique can be applied in both stages of the top-down human pose estimation approaches. For human detections, there are two NMS methods: standard NMS and soft-NMS[44]. The soft-NMS decreases the score of detecting boxes, which are suppressed in the standard NMS. The soft-NMS performs better in Ref. [39] while having the same computational complexity as the standard NMS, making it a simple method to improve human detection. Part-based NMS[11,36,39,45–48] can be performed to eliminate the reductant skeleton instance in the same detection box. The method proposed in Ref. [46] merges the parts across both time and space by substituting the medoids with centroids in the standard NMS[11]. However, this method is only designed for the match stick models. A past study[39] proposed an OKS-

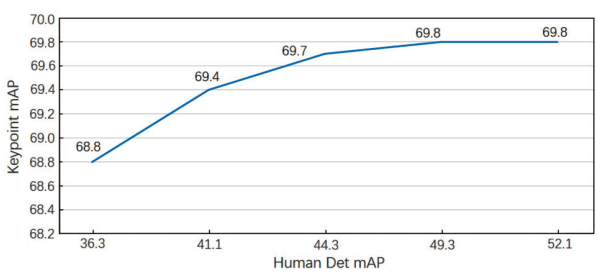


Fig. 9 Relationship of human detection mAP and keypoints mAP in Ref. [39]. The data of this chart is from the slides presented by author.

based NMS, which considers the similarity of keypoints among human instances. The parametric pose NMS proposed in Ref. [36] is data-driven, which means that all the parameters are learned from the data instead of being set manually. This method is much faster than that proposed in Ref. [46], but it has more complexity than the NMS method in Ref. [39].

3.3 Bottom-up multi-person pipeline

In contrast to the top-down approaches, the bottom-up approaches have reversed procedures, as shown in Fig. 10. All the body parts (keypoints) are detected in the first stage, then they are associated to human instances in the second stage. An illustration is given in Fig. 11. As can be seen, the inference time for the bottom-up methods is likely to be faster because it does not need to detect the pose for each person separately.

Pishchulin et al.[9] proposed the first deep learning based bottom-up pose estimation approach. Their formulation process performs the estimation task by solving a minimum-cost multi-cut problem, which models part candidates as vertices and relations between part candidates as edges.

Insafutdinov et al.[2] improved Ref. [9] in three aspects: (1) The network they used is deeper than the one in Ref. [9], which generates more effective body parts proposals. (2) The novel image-conditioned pairwise terms make it possible to assemble the part proposals into a variable number of human instances. (3) The optimization strategy is changed, which leads to both better accuracy and faster speed. Although significant improvements have been achieved compared with the previous version[9], deepercut[2] is still slow when it comes to solving the minimum cost, multi-cut problem. The method proposed in Ref. [49] further improved the method proposed in Ref. [2] by simplifying the body-part relationship graph and offloading a substantial share of computation onto a feed-forward network.

Cao et al.[3] proposed an effective method, which uses non-parametric representation called Part Affinity Fields (PAFs). After the heatmaps and PAFs are generated, a greedy algorithm is exploited to generate the person instance. Zhu et al.[50] have conducted several modifications of Ref. [3] to achieve better results. The modifications include a deeper base network and reductant PAFs, which help connect these child connections to a broken parent link.

Meanwhile, associative embedding, a method for supervising convolutional neural networks for the task of detection and grouping, is proposed in Ref. [51]. This approach simultaneously predicts part heatmaps and tagging heatmaps. The values in tagging heatmaps are similar for the same person while they are dissimilar for different persons.

3.4 Discussion for the bottom-up pipeline

3.4.1 How are heatmaps generated for the bottom-up pipeline?

The positions of keypoints are the first thing to consider

Fig. 10 Framework of bottom-up pipeline.

(a) (b) (c)

Fig. 11 An illustration of bottom-up pipeline. (a) Input image, (b) keypoints of all the person, and (c) all detected keypoints are connected to form human instance.

in human pose estimation because their locations significantly affect the performance of algorithms. There are three ways to generate ground truth heatmaps in current works. The first one is to set the heatmaps with the 2D Gaussian activation at each key point location[3]. The second way is to set the value of pixel to one in all the position within the circle, whose center is the keypoint and the radius is R (a hyper parameter), while other positions are set to zero[1,2,9]. When this type of heatmap is employed, position offset maps are predicted to locate the keypoints more accurately. The third way is to generate one-hot binary mask, where only a single pixel is labeled as the foreground. The softmax loss is used when one-hot heatmaps are employed as they encourage a single point to be detected[37]. The max activations of heatmaps (or the heatmaps performed with the Gaussian filter) have been selected as keypoints in some works[3,4]. Another study[1] also used heatmaps to vote for the final positions of keypoints.

3.4.2 Comparison of the classical methods and deep learning methods for point association

The association of detections is an important step in the bottom-up approaches. Deepcut[9] uses CNN just to learn the appearance features, using other manually defined geometric features to fit the logistic model for pairwise probability estimation. However, Deepercut[2] changes the manually calculated features to learned features generated by the deep neural network, which improves the AP by a large margin. Both of them apply the logistic model for geometric features in order to model the pairwise joints affinity. Both PAFs[3,50] and associative embedding[51] are learned simultaneously with heatmaps in deep learning fashion. They are more direct when it comes to grouping joints to human instances. The performance of these two approaches is better than that reported in Ref. [9]. This is because deep neural networks have larger capacity and are learned from data directly, which can capture both the local features and the global context.

4 Discussion

The pipeline and related approaches have been reviewed above, but the common features and differences among works are not discussed in detail. In this section, some key procedures (listed in Table 2) are discussed to reveal what matters in both two pipelines.

4.1 How is data augmentation performed?

The performance of deep learning methods highly relies on the data input. The more data a machine learning model can gain access, the more effective the model is. Most previous works utilized data augmentation to enhance the generalization of neural networks. Simple but effective data augmentation techniques for human pose estimation include cropping, rotating, scaling, and horizontally flipping input images. There are other promising deep learning methods as well. Using unlabeled data to train a network is an underdeveloped direction. Data distillation[35], an omni-supervised learning method, utilizes unlabeled data for data augmentation. This work proves that both similar and dissimilar unlabeled data are useful and effective for pose estimation.

4.2 What is the common method for data preprocessing?

Training data in datasets have variant sizes, which makes it difficult for a network to learn from non-uniform data. To make the inputs uniform, the images are always resized during the training process. There are two important factors for image resizing: ratio and absolute size. In the human pose estimation task, each

Table 2 A list of key procedures.

Key procedure	Approach	Single	Top-down	Bottom-up
Data augmentation	Traditional: cropping, rotating, scaling, and horizontally flipping	✓	✓	✓
	Using unlabeled data: Data distillation	✓	✓	
Data preprocessing	Resize without distortion	✓	✓	✓
Network design	Hole algorithm, Upsampling, Output stride <= 8, Skip connections, Big Effective receptive field, Search automatically	✓	✓	✓
Post-processing	Detection NMS		✓	
	Skeleton NMS		✓	✓

human detection box is extended to a fixed ratio without distorting the image, because the data distribution is kept in this way. A bigger training image can always produce gain for the network. A possible explanation for this is that a bigger input has better resolution, which provides more detailed information for the network. The ratio of input size also affects the training result. Normally, the width and height of the resized images are set to be the same. However, Ref. [39] reported that the input image with shorter width shows minor decrease of AP because most of the persons in the dataset have a shorter width. Bigger input images occupy more memory in training stage while it can bring better gain.

4.3 Comparison of different network structures

Network design is the fundamental work for deep learning. There are numerous network structures[40–42, 52, 53] proposed for classification task in recent years. These structures are then transferred to other tasks, such as object detection, image segmentation, and human pose estimation. However, it is inappropriate to apply these networks directly to human pose estimation because a higher resolution feature map is required to obtain a more accurate result. To solve this problem, two structures are designed to perform the required process. The first one is hole algorithm (also called atrous convolution or dilation convolution)[54], which enables the layers to increase the receptive fields effectively while keeping the output resolution fixed. The second one is upsampling, which increases the special resolution of the input feature map. The latter one shows better performance in segmentation task[2], but it occupies more GPU memory. Both these techniques are employed in Refs. [1, 2] to obtain an appropriate feature map size.

The stride of output is another factor to consider because a small stride usually consumes more storage with the same depth but outputs higher-resolution heatmaps. The final output stride is 8 in most of the previous works. In Ref. [4], stride 8 networks have been shown to achieve similar results as stride 4 networks. However, Ref. [4] is for single-person pose estimation, where the detections are resized before being sent to the network, so whether a smaller stride is required is still under exploration.

Moreover, compared to the classification and detection task, human pose estimation is more sensitive to the location of keypoints. To this end, skip connections between shallow layers and deep layers are used in many works[4, 39, 49]. Shallow layers contain more local information due to their high resolution, whereas deep layers contain more semantic information. Skip connections benefit this task by combining the local and semantic information.

The size of the effective receptive field is also a key factor in designing a good network. The receptive field should be big enough to cover the whole body of a person so all the context information can be included for human keypoint detection[4, 49].

In addition, Zhong et al.[55] proposed a method to design high-performance network blocks automatically with Q-learning, which is applied in Ref. [45].

4.4 Bottom up or top down?

As presented above, both the bottom-up and top-down approaches have been explored using deep learning methods in recent years. However, which approach is better than the other has yet to be identified because multiple aspects are considered in real-world applications. Accuracy and speed are two crucial factors for multi-person pose estimation evaluation.

Accuracy: Both the winner[3] of the COCO 2017 keypoint and the winner[39] of the AI Challenger Human Skeletal System Keypoints Challenge employed the top-down pipeline. However, Ref. [51] reported that the bottom-up pipeline can achieve similar accuracy while multiscale results are fused for testing. To consider the bottom-up pipeline, the scale variety of persons may bring difficulties for human pose estimation because the network cannot learn consistent features from the images. Another fact that should be noted is that output feature map resolution and network capacity are constrained by GPU capacity. The average resolution of a single person in a bottom-up pipeline is lower than that in the top-down pipeline during the training stage with the same network and same GPU storage. Hence, what really constrains the accuracy of the bottom-up pipeline may be the hardware limit.

Speed: The pose of each person in the top-down pipeline is estimated one by one, which consumes linear time with the increase of human number. In contrast, the image only goes through the network once in the bottom-up pipeline. It always takes more time when the image does forward propagation as current networks are always very deep. However, the grouping time can be very short if the method is designed properly (In Ref. [3], the grouping time for 9 persons is 0.58 ms). Therefore, faster speed is possibly achievable in the

bottom-up pipeline.

4.5 Are there other tricks for human pose estimation?

Most of techniques are illustrated above, but there are still other tricks employed in recent works. For example, the intermediate supervision used in Refs. [4, 39, 49] has been proven to be effective in training a better network because it forces the network to learn features from shallower layers and prevents the vanishing gradient.

Multi-task learning improves the model performance in detection tasks[56] and keypoint detection tasks[37, 56], because the network can learn more robust feature representations when the tasks are similar.

Another trick is only for the bottom-up pipelines. Single-person keypoint detection models are used to improve the bottom-up pipelines[3] and it has demonstrated a huge improvement in accuracy although it slows down the detection speed.

5 Metrics and Datasets

5.1 Dataset

Human pose estimation has been studied for years. However, it is difficult to create a universal dataset for this task because human poses are variant. To solve the estimation problem step by step, the trends of both number and complexity of datasets are increasing. The list of human pose datasets is shown in Table 3 and some samples of these datasets are shown in Fig. 12. Early datasets[24] contain images with relatively simple backgrounds. However, deep learning based methods are not suitable for these datasets because the number of images is too small for training. The common datasets used in deep learning based approaches include MSCOCO[61], MPII[59], LSP[57], FLIC[34], PoseTrack[63], and AI Challenger[62], which

Table 3 Human pose estimation datasets.

Year	Dataset	Size	Type	Num	Description
2008	Buffy[24]	472 frames training 276 frames testing	Upper body	6 body parts	Data are from TV show. Line segments are provided to indicate position. Size and orientation of body parts are also provided. Only one person is annotated in each image.
2010	LSP[57]	1000 images training 1000 images testing	Full body	14 keypoints	Data are from Flicker with sport category tag. Images are scaled. Only one person is annotated in each image.
2013	FLIC[34]	3987 images training 1016 images testing	Upper body	10 keypoints	Data are from Hollywood movies. The persons are occluded or severely non-frontal are deleted.
2014	Parse[58]	100 images training 205 images testing	Full body	14 keypoints	It is a small dataset with extended annotations including facial expression, gaze direction, and gender.
2014	MPII Human pose[59]	410 activities 2.5×10^4 images	Full body	16 keypoints	Data are from YouTube videos. It covers 410 human activities and each image is provided with activity label.
2014	Poses in the wild[60]	30 sequences 900 frames	Upper body	5 keypoints	The data are 30 videos sequences generated from 3 Hollywood movies.
2014	MSCOCO[61]	115×10^3 images training 5×10^3 images validation 20×10^3 images test-Dev 20×10^3 images test-Challenge	Full body	17 keypoints	Data are from Internet. It contains diverse activities.
2017	AI Challenger[62]	210×10^3 images training 30×10^3 images validation 60×10^3 images testing	Full body	14 keypoints	Data are crawled from Internet. It is the largest human pose image dataset currently.
2017	PoseTrack[63]	514 videos including 66 374 frames 300 videos training 50 videos validation 208 videos testing	Full body	15 keypoints	The videos are from MPII Human Pose dataset. This dataset focusses on 3 aspects: (1) single-frame multi-person pose estimation. (2) multi-person pose estimation in videos. (3) multi-person articulated tracking.

Fig. 12 Examples of different datasets.

contain more images in more complicated scenes. The LSP and FLIC datasets are relatively small and only contain specific categories of activity. The images in the LSP dataset are from a sports scene. The FLIC datasets are collected from Hollywood movies. The latest datasets, such as MSCOCO and AI Challenger, are bigger in both size and number of categories.

5.2 Metrics

Evaluating the performance of human pose estimation is difficult, because many factors need to be considered. An evaluating metric used in early works is the Percentage of Correctly estimated body Parts (PCP)[24], which evaluates stick predictions. Another widely used metric for keypoint detection is PCK[11] and its variant, PCKh. In these two metrics, a point is correct if it falls within the $\alpha \cdot \max(w, h)$ pixels of the ground truth keypoint, where w and h are the height and width of the bounding box of the person (head of the person in PCKh), respectively. Other recent metrics include the Object Keypoint Similarity (OKS) and AP of OKS[64], which not only take scale into consideration, but also introduce the per-point constant to control falloff.

6 Conclusion

In this survey, we presented a comprehensive review for deep learning based human pose estimation method. Although the current human pose estimation methods have been improved significantly, they can still be improved for better real-world applications. The speed of the current algorithms is still slow and cannot meet the requirements of real-time prediction, so accelerating the detection speed must be further explored. There are already some works that investigated network compression and network accelerating, but they are not designed for human pose detection, which needs higher resolution output feature maps compared with the classification task and detection task. Therefore, the performance of existing algorithms is still under verification. Accelerating methods should be further

explored.

The current dataset is very large, but the pose distribution is unbalanced, and no study has explored ways to detect rare poses with an unbalanced dataset. The possible improvements include doing data augmentation and designing a special training procedure. GANs[28] and the employment of unlabeled data are two aspects of data augmentation.

Occlusion and self-occlusion still pose challenges for human pose estimation. Some works combine human priors and data-driven methods to solve the problem, but their results are not robust enough. Human priors are still under improvement before they can fully yield a satisfying performance.

Our taxonomy is based on methodology which includes single-person pipeline and multi-person pipeline. The comparisons are made among different frameworks and different pipelines. Some key procedures are discussed which include data augmentation, data preprocessing, network design, and post-processing. We also summarized dataset and metrics for deep learning based human pose estimation. Hopefully, readers can get inspired from our survey and solve the difficulties we mentioned above.

We hope our survey can motivate new research efforts to advance the field of research on estimation speed accelerating, unbalanced and unlabeled data based data augmentation, and occlusion resisted pose estimation.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 61673192, 61573219, and 61472163), the Fund for Outstanding Youth of Shandong Provincial High School (No. ZR2016JL023), the National High-Tech Research and Development Plan (No. 2015AA042306), and the National Social Science Fund Project (No. 13CTQ010).

References

[1] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, Towards accurate multiperson pose estimation in the wild, arXiv preprint arXiv:1701.01779, 2017.

[2] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, Deepercut: A deeper, stronger, and faster multi-person pose estimation model, in *European Conference on Computer Vision*, 2016, pp. 34–50.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in *CVPR*, 2017, vol. 1, p. 7.

[4] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, Convolutional pose machines, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4724–4732.

[5] A. Toshev and C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, WI, USA, 2014, pp. 1653–1660.

[6] X. Chen and A. L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, in *Advances in Neural Information Processing Systems*, 2014, pp. 1736–1744.

[7] J. Martinez, R. Hossain, J. Romero, and J. J. Little, A simple yet effective baseline for 3d human pose estimation, in *IEEE International Conference on Computer Vision*, Venice, Italy, 2017, vol. 206, p. 3.

[8] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, Adversarial posenet: A structure-aware convolutional network for human pose estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1212–1221.

[9] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, Deepcut: Joint subset partition and labeling for multi person pose estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4929–4937.

[10] Y. Yang and D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in *Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 1385–1392.

[11] Y. Yang and D. Ramanan, Articulated human detection with flexible mixtures of parts, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[12] F. Wang and Y. Li, Beyond physical connections: Tree models in human pose estimation, in *Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, 2013, pp. 596–603.

[13] M. Sun and S. Savarese, Articulated part-based model for joint object detection and pose estimation, in *Computer Vision (ICCV)*, Barcelona, Spain, 2011, pp. 723–730.

[14] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, 2d articulated human pose estimation and retrieval in (almost) unconstrained still images, *International Journal of Computer Vision*, vol. 99, no. 2, pp. 190–214, 2012.

[15] M. Eichner and V. Ferrari, We are family: Joint pose estimation of multiple persons, in *European Conference on Computer Vision*, Crete, Greece, 2010, pp. 228–242.

[16] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, Deep learning for visual understanding: A review, *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[17] R. Poppe, Vision-based human motion analysis: An overview, *Computer Vision and Image Understanding*, vol. 108, nos. 1&2, pp. 4–18, 2007.

[18] Z. Liu, J. Zhu, J. Bu, and C. Chen, A survey of human pose estimation: The body parts parsing based methods,

Journal of Visual Communication and Image Representation, vol. 32, pp. 10–19, 2015.

[19] H.-B. Zhang, Q. Lei, B.-N. Zhong, J.-X. Du, and J. Peng, A survey on human pose estimation, *Intelligent Automation & Soft Computing*, vol. 22, no. 3, pp. 483–489, 2016.

[20] W. Gong, X. Zhang, J. Gonźalez, A. Sobral, T. Bouwmans, C. Tu, and E.-h. Zahzah, Human pose estimation from monocular images: A comprehensive survey, *Sensors*, vol. 16, no. 12, p. 1966, 2016.

[21] E. Murphy-Chutorian and M. M. Trivedi, Head pose estimation in computer vision: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[22] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, Vision-based hand pose estimation: A review, *Computer Vision and Image Understanding*, vol. 108, nos. 1&2, pp. 52–73, 2007.

[23] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, A survey on deep learning based approaches for action and gesture recognition in image sequences, in *Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA, 2017, pp. 476–483.

[24] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, Progressive search space reduction for human pose estimation, in *Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[25] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, Human pose estimation with iterative error feedback, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4733–4742.

[26] X. Sun, J. Shang, S. Liang, and Y. Wei, Compositional human pose regression, in *The IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, vol. 2.

[27] D. C. Luvizon, H. Tabia, and D. Picard, Human pose regression by combining indirect part detection and contextual information, arXiv preprint arXiv:1710.02322, 2017.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 2672–2680.

[29] A. Newell, K. Yang, and J. Deng, Stacked hourglass networks for human pose estimation, in *European Conference on Computer Vision*, Amsterdam, Netherlands, 2016, pp. 483–499.

[30] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, Multi-context attention for human pose estimation, arXiv preprint arXiv:1702.07432, 2017.

[31] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman, Deep convolutional neural networks for efficient pose estimation in gesture videos, in *Asian Conference on Computer Vision*, Singapore, 2014, pp. 538–552.

[32] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in *Advances in Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 1799–1807.

[33] T. Pfister, J. Charles, and A. Zisserman, Flowing convnets for human pose estimation in videos, in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1913–1921.

[34] B. Sapp and B. Taskar, Modec: Multimodal decomposable models for human pose estimation, in *Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, 2013, pp. 3674–3681.

[35] I. Radosavovic, P. Dollar, R. Girshick, G. Gkioxari, and K. He, Data distillation: Towards omni-supervised learning, arXiv preprint arXiv:1712.04440, 2017.

[36] H. Fang, S. Xie, Y.-W. Tai, and C. Lu, Rmpe: Regional multi-person pose estimation, in *The IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[37] K. He, G. Gkioxari, P. Dollar, and R. Girshick, Mask rcnn, in *Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980–2988.

[38] U. Iqbal and J. Gall, Multi-person pose estimation with local joint-to-person associations, in *European Conference on Computer Vision*, Amsterdam, Netherlands, 2016, pp. 627–642.

[39] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, Cascaded pyramid network for multi-person pose estimation, arXiv preprint arXiv:1711.07319, 2017.

[40] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.

[41] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in *AAAI*, San Francisco, CA, USA, 2017, vol. 4, p. 12.

[43] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, in *CVPR*, Honolulu, HI, USA, 2017, vol. 1, p. 4.

[44] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, Improving object detection with one line of code, arXiv preprint arXiv:1704.04503, 2017.

[45] Y. W. Wang, C. Wang, Q. Li, B. Leng, Z. Li, and J. Yan, Team oks keypoint detection, http://presentations.cocodataset.org/COCO17-Keypoints-TeamOKS.pdf, 2017.

[46] X. P. Burgos-Artizzu, D. C. Hall, P. Perona, and P. Dollár, Merging pose estimates across space and time, in *British Machine Vision Conference (BMVC),* Bristol, UK, 2013.

[47] X. Chen and A. Yuille, Parsing occluded people by flexible compositions, in *Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 3945–3954.

[48] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, Detect what you can: Detecting and representing objects using holistic models and body parts, in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, WI, USA, 2014, pp. 1971–1978.

[49] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, Arttrack: Articulated multi-person tracking in the wild, in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Honolulu, HI, USA, 2017.

[50] X. Zhu, Y. Jiang, and Z. Luo, Multi-person pose estimation for posetrack with enhanced part affinity fields, presented at the ICCV PoseTrack Workshop, Venice, Italy, 2017.

[51] A. Newell, Z. Huang, and J. Deng, Associative embedding: End-to-end learning for joint detection and grouping, in *Advances in Neural Information Processing Systems*, San Francisco, CA, USA, 2017, pp. 2274–2284.

[52] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, Aggregated residual transformations for deep neural networks, in *Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 5987–5995.

[53] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, vol. 1, p. 3.

[54] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[55] Z. Zhong, J. Yan, and C.-L. Liu, Practical network blocks design with q-learning, arXiv preprint arXiv:1708.05552, 2017.

[56] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in *Advances in Neural Information Processing Systems*, Austin, TX, USA, 2015, pp. 91–99.

[57] S. Johnson and M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, in *Proceedings of the British Machine Vision Conference*, Aberystwyth, UK, 2010.

[58] S. Antol, C. L. Zitnick, and D. Parikh, Zero-shot learning via visual abstraction, in *European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 401–416.

[59] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, WI, USA, 2014, pp. 3686–3693.

[60] A. Cherian, J. Mairal, K. Alahari, and C. Schmid, Mixing body-part sequences for human pose estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, WI, USA, 2014, pp. 2353–2360.

[61] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, Microsoft coco: Common objects in context, in *European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 740–755.

[62] 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 arXiv:1711.06475, 2017.

[63] M. Andriluka, U. Iqbal, A. Milan, E. Insafutdinov, L. Pishchulin, J. Gall, and B. Schiele, Posetrack: A benchmark for human pose estimation and tracking, arXiv preprint arXiv:1710.10000, 2017.

[64] Mscoco keypoint evaluation metric, http://mscoco.org/dataset/#keypoints-eval, 2017.

Qi Dang received the BS and MS degrees from Beijing University of Posts and Telecommunications, China, in 2014 and 2017, respectively. He is currently a research assistant in Tsinghua University. His research interests include computer vision and deep learning.

Jianqin Yin received the PhD and MSc degrees in pattern recognition and intelligent systems from Shandong University, China, in 2013 and 2002, respectively. From 2003 to 2016, she worked as a teacher in University of Jinan. From 2016, she works as a teacher in Beijing University of Posts and Telecommunications. Her research interests include robotics, machine learning, and computer vision.

Bin Wang received the MSc degree from Tsinghua University, China. Now he is an engineer in Tsinghua University. His research interests include robotics, deep learning, and computer vision.

Wenqing Zheng is an undergraduate in Beijing University of Posts and Telecommunications.