applied sciences

MDPI

Review

A Survey on Depth Ambiguity of 3D Human Pose Estimation

Siqi Zhang [1,†], Chaofang Wang [1,†], Wenlong Dong [1,2,*] and Bin Fan [1,2,*]

[1] Institute of Disaster and Emergency Medicine, Tianjin University, Tianjin 300072, China
[2] Wenzhou Safety (Emergency) Institute, Tianjin University, Wenzhou 325000, China
* Correspondence: dongwl@tju.edu.cn (W.D.); fanbin911@126.com (B.F.)
† These authors contributed equally to this work.

Abstract: Depth ambiguity is one of the main challenges of three-dimensional (3D) human pose estimation (HPE). The recent strategies of disambiguating have brought significant progress and remarkable breakthroughs in the field of 3D human pose estimation (3D HPE). This survey extensively reviews the causes and solutions of the depth ambiguity. The solutions are systematically classified into four categories: camera parameter constraints, temporal consistency constraints, kinematic constraints, and image cues constraints. This paper summarizes the performance comparison, challenges, main frameworks, and evaluation metrics, and discusses some promising future research directions.

Keywords: human pose estimation; depth ambiguity; deep learning; human key points positioning; human object detection

Citation: Zhang, S.; Wang, C.; Dong, W.; Fan, B. A Survey on Depth Ambiguity of 3D Human Pose Estimation. *Appl. Sci.* **2022**, *12*, 10591. https://doi.org/10.3390/app122010591

Academic Editor: Athanasios Nikolaidis

Received: 17 September 2022
Accepted: 11 October 2022
Published: 20 October 2022

1. Introduction

HPE is one of the most fundamental tasks in computer vision, which describe human pose by locating the key joint points in the input image or video [1]. In years of continuous development, it has aroused long-standing research attention. Today it plays a significant role in numerous applications such as behavior analysis, virtual reality, and human–computer interactions. Base on the dimension of estimation, it can be split into two-dimensional (2D) and 3D pose estimations [2]. Different from 2D methods locating the X and Y axis coordinates, 3D methods estimate joint position in 3D space by adding an extra axis [3]. Compared with 2D representation, 3D representation provides additional depth information. Therefore, 3D HPE contains higher research value and can be put into wider applications [4].

Recently, driven by advanced deep learning technology and large-scale datasets, 3D HPE has continually made great progress. However, the inherently existing depth ambiguity problem still seriously restricts the accuracy of 3D HPE. Due to the unknown relative depth between body joints, several 3D poses may correspond to the same 2D pose, as shown in Figure 1, which leads to depth ambiguities in the 2D to 3D projection, also known as ill-posedness. When predicting 3D human poses from static images, we are inverting an inherently lossy nonlinear transformation that combines the perspective projection and kinematics [5]. This ambiguity makes it difficult, in the absence of priors other than the joint angle limits or the body's non-self-intersection constraints, to recover the original 3D pose from its projection. The existence of monocular 3D ambiguities is well known, but it is interesting to study to what extent these are present among the poses of a large, ecological dataset. We can assess the occurrence of ambiguities by looking at 3D and 2D ground truth pose information. In contrast, since the human perception mechanism (HPM) [6], human observers could distinguish the swing of the limbs and the rotation of the joints according to information such as changes in illumination, shadows, and shapes and easily infer a reasonable 3D pose. Although there are some existing reviews for HPE, there still lacks a survey to review the cause and solutions of the depth ambiguity. The survey [7] reviewed the recent deep learning-based 2D and 3D human pose estimation

methods. This paper summarizes the performance comparison of various strategies based on the dataset Human3.6M.

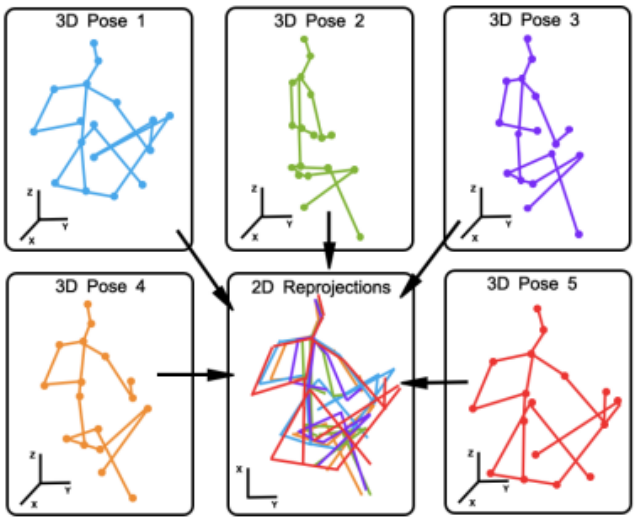

Figure 1. Illustration of depth ambiguity of 3D HPE.

The main objective of this paper is to provide a summary of disambiguating methods that address the under-constrained nature of 3D HPE from single RGB inputs. We classify the mentioned methods based on the main constrain information into four categories. Then, we analyze the performances of the mentioned methods according to their categories and discuss the advantages and the limitations of the followed strategies. In the end, we discuss the existing problems of current research and evaluate future development trends.

2. Status of Research

Human observers use two kinds of information when recognizing poses from images: appearance and constraint information. The former is the basis for locating keypoints, whereas the latter has crucial guiding significance when locating difficult keypoints. Constraint information includes the inherent mutual relationship between the keypoints, the constraint relationship formed by the interaction between the human body and the environment, etc. [8]. Currently, 3D HPE methods tend to regularize the learning process with diverse constraints to relieve ambiguity.

We evaluate various methods' performance on the motion capture dataset, Human3.6M [9]. It contains 3.6 million 3D human poses including discussion, smoking, taking photos, talking on the phone, etc., and corresponding images from four different views. For evaluation, there are three protocols with different training and testing data splits (protocol #1, protocol #2, and protocol #3) [10]. In this paper, we consider protocol #1. Protocol #1 is the mean per joint position error (MPJPE) in millimeters which is the mean Euclidean distance between predicted joint positions and ground-truth joint positions and follows.

2.1. Camera Parameter Constraints

Based on the camera imaging model, camera calibration establishes the correspondence between a point in a 3D scene and its 2D pixel position in the image [11]. Based on the basic principles of camera calibration, some researchers make good use of camera parameter constraints to alleviate ambiguity and find the coordinates in the 3D space.

2.1.1. Method Content

Habibie et al. [12] learn weak-perspective camera (WPC) parameters from a given monocular image and project the predicted 3D pose into 2D space. Such projection loss

can be used to update the position of the 3D joints. Li et al. [13] introduce the mixture density network (MDN) to generate multiple alternative 3D poses that mimic the depth ambiguity problem. Based on the WPC model, they compared the 2D re-projection of multiple 3D pose hypotheses to select the 3D pose with the highest similarity. The WPC model-based approach mentioned above may greatly simplify the imaging model, which can easily lead to the failure of some pose estimations. To address this issue, as shown in Figure 2, Moon et al. [14] designed a camera distance-aware model to obtain the absolute depth from the camera to the target, allowing the RootNet to adapt to various camera parameters flexibly. The existing unsupervised framework, generative adversarial network (GAN) [15], has excellent generalization capabilities. Wandt et al. [16] exploit a GAN-based architecture to predict camera parameters from the input image and regard the predicted camera parameters as a weak-perspective projection matrix. The model performs weakly supervised learning by globally scaling the estimated 3D pose and re-projecting it into 2D space to determine the deviation of the predicted pose from the original 2D input.

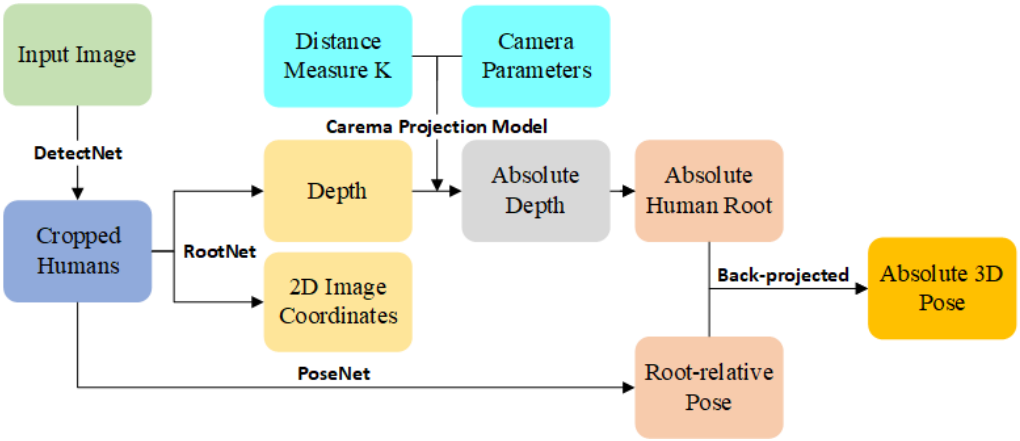

Figure 2. Illustration of Moon's camera distance-aware top-down approach.

2.1.2. Performance Analysis

As shown in Table 1, because most poses of the Human3.6 dataset are standing poses and motions, e.g., sit or sit down, contain more ambiguities and occlusions, it provides sufficient proof of the disambiguating effectiveness in Moon et al. [14]. Although it does not use any ground truth information in the estimation process, it still achieves low errors, especially for ambiguous motions such as photo and walk together. In contrast, the disambiguating effects are limited in Habibie et al. [12] and Li et al. [13], which use the weak-perspective model. For a reasonable comparison with other supervised methods, the unsupervised method of Wandt et al. [16] is compared with 2D groud-truth annotations and even outperforms some supervised methods.

Table 1. Summary of performance on Human3.6 Protocol #1 (MPJPE: mm).

Methods	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Habibie et al. [12]	54	65.1	58.5	62.9	67.9	75	54	60.6	82.7	98.2	63.3	61.2	66.9	50	56.5	65.7
Li et al. [13]	43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62	73.4	54.8	50.6	56	43.4	45.5	52.7
Moon et al. [14]	31	30.6	39.9	35.5	34.8	37.6	30.2	32.1	35	43.8	35.7	30.1	35.7	24.6	29.3	34
Wandt et al. [16]	33.6	38.8	32.6	37.5	36	44.1	37.8	34.9	39.2	52	37.5	39.8	40.3	34.1	34.9	38.2

2.2. Temporal Consistency Constraints

Because the depth information contained in a single image is limited, there must be multiple results when the human body is projected from 2D to 3D by using a neural network composed of almost entirely convolutional layers. The network can obtain richer depth information from the sequence. Therefore, some researchers use the video frame sequence, which is data with time information, as a dimensional supplement. The context

information provided by the adjacent frame sequence can assist in predicting the pose of the current frame.

2.2.1. Method Content

Hossain et al. [17] address the interframe incoherence and independence by propagating joint position information between frames. The model introduces temporal smoothness loss to constrain the temporal consistency, preventing excessive changes between two frames. Their skip connection learns the difference between 3D poses in different moments, making it easier to disambiguate. Due to the advantages of processing multiple sequences in parallel, lower computational complexity, and fewer model parameters, the temporal convolutional network (TCN) is often used to process coherent sports information in image sequences. Pavllo et al. [10] employ TCN to conduct dilated temporal convolutions capturing long-term information, as shown in Figure 3. Cai et al. [18] solve ambiguity by proposing a spatiotemporal graph convolutional networks (ST-GCN) to combine spatial configurations and temporal consistencies. Zhang et al. [19] introduce dynamic spatial graph (DSG) and dynamic temporal graph (DTG) convolution to calculate the spatiotemporal relationship between human joints. Their dynamical graph network can identify the key points with motion consistency to reduce the ambiguity when lifting a 2D pose to a 3D pose. Due to the fact that human bone lengths remain consistent across video, Chen et al. [20] predict the bone lengths in a specific frame of a video to alleviate the ambiguity of 3D pose estimation. In addition, the visibility scores of 2D key points are utilized in this method as additional knowledge.

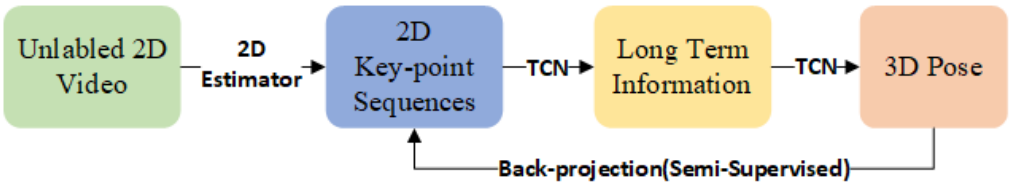

Figure 3. Illustration of Pavllo's fully convolutional architecture.

2.2.2. Performance Analysis

As shown in Table 2, thanks to temporal information, Pavllo's [10] model reduces the error by about 5 mm on average compared to its single-frame baseline model (where the width of all convolution kernels was set to W = 1). The gap between them is more significant on some more dynamic motions, such as walk (−6.7 mm) and walk together (−8.8 mm). By using the same 2D detector with Pavllo et al. [10], Chen et al. [20] obtain a more smooth prediction with lower MPJPE and achieve better performance on some difficult poses, such as sit (−3.4 mm) and sit down (−5.6 mm), which can be attributed to accurate prediction of bone length for these poses. In Chen et al. [20], the joint displacement loss can effectively guide the model to predict high-quality bone orientation according to the predicted bone length even if the body is bent. The approach designed by Zhang et al. [19] achieves lower errors on complex motions involving high ambiguities and motion uncertainty, such as sit down and photo, which shows that their method adaptively learns joint spatiotemporal relationships to capture human poses in different motions and has a good effect on disambiguation.

Table 2. Summary of performance on Human3.6 Protocol #1 (MPJPE: mm).

Methods	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Hossain et al. [17]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Pavllo et al. [10], single-frame	47.1	50.6	49	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
Pavllo et al. [10], 243-frames	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44	49	32.8	33.9	46.8
Cai et al. [18]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Zhang et al. [19]	28.5	33.5	28.1	28.9	32.6	35.5	33.3	30	37.4	39.9	31.4	30.2	29.5	23.9	25.5	31.2
Chen et al. [20]	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46	31.5	32.7	44.1

2.3. Kinematic Constraints

Kinematic constraints refer to the restrictions that the human body should obey when moving, such as body structural connectivity constraints, distance constraints between joints, the range constraints of joint angles, and the constraints that each part of the body cannot penetrate each other. These constraints can be used as hard constraints to divide the state space into legal and illegal parts to reduce the search range [21] and can also be used as soft constraints, namely penalty factors [22].

2.3.1. Method Content

Since the Euclidean distances matrix (EDM) is coordinate-free and invariant to in-plane image rotations and translations, Moreno et al. [23] use EDM to represent pairwise distances of body joints. Combining body structural information and capturing joint correlations, they formulate the 3D estimation task as a regression between matrices encoding 2D and 3D joint distances to reduce ambiguity. As shown in Figure 4, Lee et al. [24] employ propagating LSTM networks (p-LSTMs) connected in series to elaborate the 3D pose progressively. They effectively overcome ambiguity by learning joint interdependency based on actual human behavior. Wang et al. [25] propose to distinguish joints with different degrees of freedom (DOF) based on prior knowledge about physiological structure. In detail, the bi-directional dependencies among body parts with different DOFs make them supervise each other, yielding physically constrained and plausible pose-estimation results. Wang et al. [26] propose an attention mechanism-based [27] model to combine the end-to-end training scheme in GNN and the limb length constraints in PSM. In addition, Angjoo et al. [28] parameterize the human body by shape, and joint angles based on the SMPL model [29] to reduce the probability of generating an unreasonable body caused by ambiguity. As shown in Figure 5, Xu et al. [30] split the 3D coordinate regression problem into length and direction estimation, which alleviates ambiguity and self-focusing contained in images by a large margin within a more compact space.

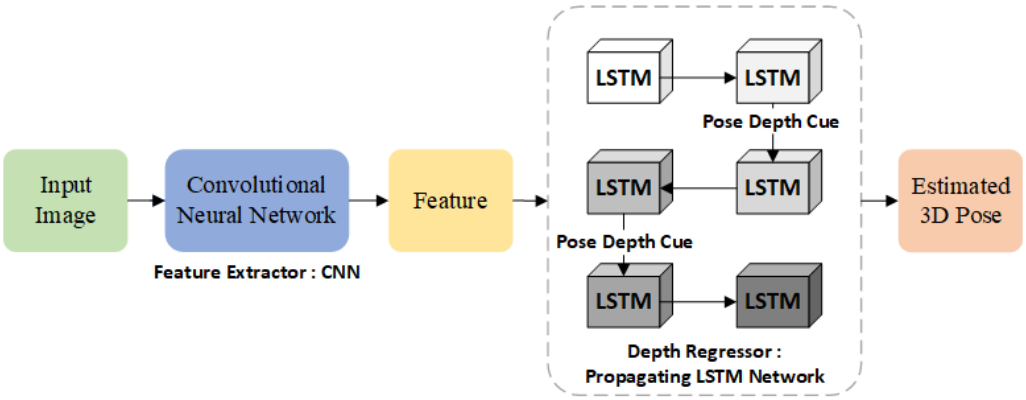

Figure 4. Illustration of Lee's approach with pose depth cues.

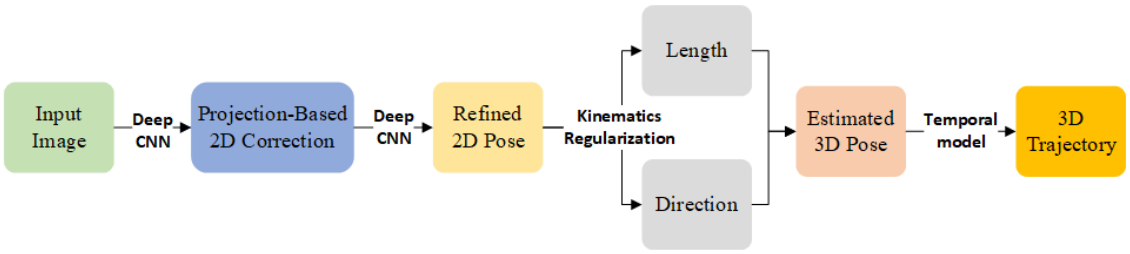

Figure 5. Illustration of Xu's approach based on kinematics analysis.

2.3.2. Performance Analysis

Explicitly incorporating kinematic analysis into models and maintaining reasonable spatiotemporal structure and compact output space, as shown in Table 3, methods (Lee et al. [24]; Wang et al. [25]; Wang et al. [26]; Xu et al. [30]) show good performance and finally improve the estimation accuracy. In particular, the methods of Wang et al. [26] and Xu et al. [30] both obtain stable performance on several motions with relatively large ambiguity problems (such as sit, sit down, and photo). However, the model of Wang et al. [26] is still challenging on the task of estimating overall attention when multiple joints are occluded. There are several defects in Angjoo et al. [28]. (1) It is difficult for the model to directly regress the internal parameters; (2) In the case of a small amount of data, it is difficult for the model to use GAN for supervision to obtain robust results, and it may even be possible worse. (3) The model does not supervise the human mesh part, which may waste resources.

Table 3. Summary of performance on Human3.6 Protocol #1 (MPJPE: mm).

Methods	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Moreno et al. [23]	69.5	80.2	78.2	87.0	100.8	102.7	76.0	69.7	104.7	113.9	89.7	98.5	82.4	79.2	77.2	87.3
Lee et al. [24]	43.8	51.7	48.8	53.1	52.2	74.9	52.7	44.6	56.9	74.3	56.7	66.4	68.4	47.5	45.6	55.8
Wang et al. [25]	44.7	48.9	47	49	56.4	67.7	48.7	47	63	78.1	51.1	50.1	54.5	40.1	43	52.6
Wang et al. [26]	36.3	42.8	39.5	40	43.9	48.8	36.7	44	51	63.1	44.3	40.6	44.4	34.9	36.7	43.4
Xu et al. [30]	37.4	43.5	42.7	42.7	46.6	59.7	41.3	45.1	52.7	60.2	45.8	43.1	47.7	33.7	37.1	45.6

2.4. Image Cues Constraints

Existing two-stage methods usually lose image information because they assume that image cues are unavailable at the second stage. If no additional image cues are utilized, methods are generally prone to ambiguities in the 2D to 3D regression. Some researchers have proposed that taking full advantage of image cues constraints is also an effective way to mitigate depth ambiguity.

2.4.1. Method Content

Xing et al. [31] train a deep convolutional neural network (CNN) to establishe the mapping relationship between the image cues and 3D human pose codes. Their coding method is finally combined with a linear matching mechanism to construct an effective disambiguating solution. The ordinal depth refers to the relative depth between joint points rather than the absolute physical depth and is extracted from images as a clue by researchers for depth learning. Pavlakos et al. [32] employ ordinal depth annotation as weak supervision to replace 3D annotation. Their end-to-end model is not affected by the reconstruction ambiguity of the 2D detector. Sharma et al. [33] proposed a similar approach to reduce the lifting ambiguity by using the obtained ordinal depth to rank candidate 3D pose samples generated by conditional variational autoencoder (CAVE). Wang et al. [34] also use the ordinal depth information to constrain the depth between adjacent joints. As shown in Figure 6, they design a pairwise ranking CNN (PRCNN) to generate pairwise depth relationships between joints, leveraging its rich geometric features and 2D joint locations to determine an unambiguous 3D pose. Wu et al. [35] propose the limb depth

maps representation, associated depth values with image cues, to break the previous nested learning process's limit that simultaneously estimates (X, Y, Z).

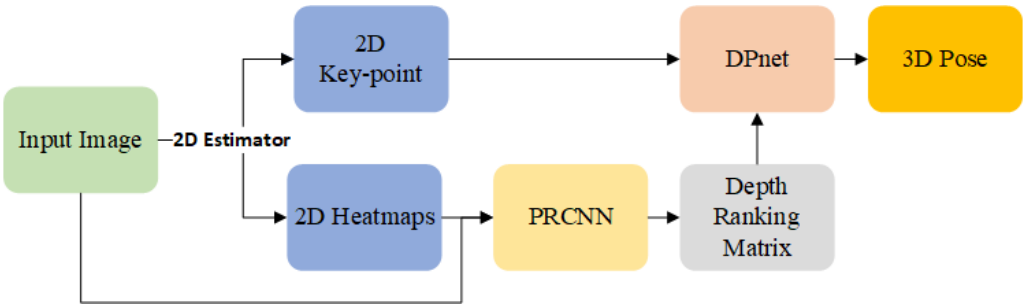

Figure 6. Illustration of Wang's two-stage approach based on depth ranking.

2.4.2. Performance Analysis

Through quantitative and qualitative experiments, Xing et al. [31] demonstrated that their method, benefiting from relieving ambiguity in 3D poses, significantly improves reconstruction accuracy, which can smoothly adapt to changes in perspective. From the data in Table 4, it can be observed that models (Pavlakos et al. [32]; Sharma et al. [33]; Wang et al. [34]) all obtain low MPJPE, showing that ordinal depth, with a special ability of geometric knowledge in resolving reconstruction ambiguity, provides weak supervision signals that can effectively enhance 3D HPE. At the same time, it prevents the network from overadapting to a specific camera perspective, alleviating the problem of definition ambiguity in the second stage of previous methods. The model of Wu et al. [35] overcomes the uncertainty of the conventional nesting method and locates the key points accurately by the depth value given explicitly. Its MPJPE value even reaches 43.2 mm, which obtains the best performance in this type of method.

Table 4. Summary of performance on Human3.6 Protocol #1 (MPJPE: mm).

Methods	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Xing et al. [31]	60.6	54.1	60.0	59.6	60.7	78.8	56.2	52.6	58.0	82.9	66.7	62.0	51.5	60.9	42.8	60.1
Pavlakos et al. [32]	48.5	54.4	54.4	52	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Sharma et al. [33]	48.6	54.5	54.2	55.7	62.6	72	50.5	54.3	70	78.3	58.1	55.4	61.4	45.2	49.7	58
Wang et al. [34]	49.2	55.5	53.6	53.4	63.8	67.7	50.2	51.9	70.3	81.5	57.7	51.5	58.6	44.6	47.2	57.8
Wu et al. [35]	34.9	40.8	37.5	47.2	41.5	46.6	35.9	39.5	52.6	72.5	42.3	45.8	42	31.6	33.8	43.2

3. Challenges

Despite the great development of disambiguating with various methods, there remain some unresolved challenges and gaps between research and practical applications. The challenges of eliminating depth ambiguity mainly have six aspects.

- Existing methods based on camera parameter constraints generally simplify the imaging model and directly treat the estimation task as a coordinate regression without adequate consideration of the inherent kinematic structure of the human body, which may lead to invalid results.
- Due to the lack of the original input image, the two-stage methods place overreliance on the 2D detector and discard the rich spatial information in the image when estimating the 3D pose. The error of the first stage will be amplified in the 3D estimation. The algorithm is ultimately limited if the 2D pose estimation is not updated with the 3D.
- The one-stage methods usually fix the scale of the 3D pose. The 3D pose is constructed from the 2D pose and depth, which may make estimation fail when the height of the subject is far from the height in the training set.

- The interframe continuity and stability constraints used by methods based on temporal consistency constraints lead to smoothness effects, which may cause inaccurate estimation of each frame. The estimated results will have floating, jitter, and sliding problems.
- Some kinematic constraints belong to simple structural constraints and are commonly treated as auxiliary losses in research, such as symmetry loss, joint angles limit, e.g., could only make marginal improvements to the estimation results, and limited disambiguating effect.
- The monocular image belongs to a two-dimensional representation and carries no depth information, which is challenging for image cue methods to learn depth information. In addition, as the depth is highly sensitive to camera parameters, translation and rotation will make joint depth prediction more difficult.

4. Future Potential Development

According to the current works and shortcomings, we propose several next future directions worthy of attention for disambiguating research as follows.

- Weakly supervised and unsupervised methods. The traditional deep convolutional neural network requires adequate manual annotations.The weakly supervised method does not need a large number of 3D annotations but 2D annotated data does, which reduces algorithm costs. Unsupervised methods follow this trend even more so.
- Interaction and reconstruction between scene object and human. The contact between human and object is the essential visual cue for inferring 3D movement. The current works based on a single RGB image generally first identify a human target using a bounding box and then estimate the cropped body, rarely paying attention to the scenes and objects that contain rich clues. The 3D HPE can utilize interpenetration constraints to limit the intersection between the body and the surrounding 3D scene.
- The 3D HPE from multi-view images. Multi-view images can significantly mitigate ambiguity, and their typical methods include fusing multi-view 2D heatmaps, enforcing consistency constraints between multiple views, and triangulation measurement, etc.
- Sensor technology. Some works use sensors, such as RGB-D depth cameras, inertial measurements units (IMUs) and radio frequency (RF) to add the collected depth, joint direction and other information [36–38]. Compared with depth images, using sensors to capture point clouds can provide more information.

5. Conclusions

Human pose estimation is a hot research area in computer vision that evolved recently along with the blooming of deep learning. Although commercial products such as Kinect with depth sensor, and TheCaptury with multiple cameras have been employed for 3D body pose estimation, all these systems work in very constrained environments or need special markers on the human body. A monocular camera, as the most widely used sensor, is very important for 3D human pose estimation. Deep neural networks have the capability to estimate the dense depth and sparse depth points (joints) as well from monocular images. Due to limitations in depth ambiguity problem, early networks were inaccurate on 3D HPE. This review summarizes recent progress in disambiguation schemes from monocular RGB images or videos. However, the overreliance on the 2D detector remains extremely challenging. As for the case of camera parameter constraints, camera perspective models are less developed. We point out that kinematic and image cue constraints that clearly work are still far from being established. Most recently, weak supervision and unsupervision have drawn significant attention. Furthermore, multi-view images effectively solve this problem, so we can expect many innovations in the next few years, especially when sensor technologies are applied to this field. In addition, we conjecture that research on interaction and reconstruction between scene objects and the human body will also be a promising direction in the future.

Author Contributions: Conceptualization, S.Z. and C.W.; methodology, S.Z.; software, C.W.; validation, S.Z. and C.W.; formal analysis, S.Z.; investigation, S.Z.; resources, B.F.; data curation, C.W.; writing—original draft preparation, S.Z.; writing—review and editing, S.Z. and C.W.; supervision, W.D. and B.F.; project administration, W.D. and B.F.; funding acquisition, W.D. and B.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "Research on artificial intelligence cardiopulmonary resuscitation training and assessment system, TJUWYY2022014", the "National key research and development plan project of China, 2018YFC1504405-3" and the "Tianjin University Independent Fund Project, 2020XZS-0029".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Y. Research progress of two-dimensional human pose estimation based on deep learning. *Comput. Eng.* **2021**, *47*, 1–16.
2. Han, G.J. A survey of two dimension pose estimation. *J. Xi'an Univ. Posts Telecommun.* **2017**, *22*, 1–9.
3. Gamra, M.B. A review of deep learning techniques for 2D and 3D human pose estimation. *Image Vis. Comput.* **2021**, *114*, 104282. [CrossRef]
4. Wang, J. Deep 3D human pose estimation: A review. *Comput. Vis. Image Underst.* **2021**, *210*, 103225–103246. [CrossRef]
5. Kocabas, M.; Athanasiou, N.; Black, M. VIBE: Video Inference for Human Body Pose and Shape Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5252–5262.
6. Yang, W.; Ouyang, W.; Wang, X. 3D human pose estimation in the wild by adversarial learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5255–5264.
7. Chen, Y. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [CrossRef]
8. AID: Pushing the Performance Boundary of Human Pose Estimation with Information Dropping Augmentation. Available online: https://arxiv.org/abs/2008.07139 (accessed on 23 July 2022).
9. Ionescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339. [CrossRef] [PubMed]
10. Pavllo, D.; Feichtenhofer, C.; Grangier, D. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7753–7762.
11. Heng, W. Calibration and Rapid Optimizing of Imaging Model for a Two-camera Vision System. *J. App. Sci.* **2002**, *20*, 225–229.
12. Habibie, I.; Xu, W.; Mehta, D.; Pons-Moll, G.; Theobalt, C. In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10897–10906.
13. Li, C.; Lee, G.H. Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9879–9887.
14. Moon, G.; Chang, J.Y.; Lee, K.M. Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 10132–10141.
15. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
16. Wandt, B.; Rosenhahn, J. RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7774–7783.
17. Hossain, M.; Little, J.J. Exploiting temporal information for 3D pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 69–86.
18. Cai, Y.J.; Ge, L.H.; Liu, J.; Cai, J.F.; Cham, T.J.; Yuan, J.S.; Thalmann, N.M. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2272-–2281.
19. Zhang, J.H. Learning Dynamical Human-Joint Affinity for 3D Pose Estimation in Videos. *IEEE Trans. Image Process.* **2021**, *30*, 7914–7925. [CrossRef] [PubMed]
20. Chen, T.L. Anatomy-aware 3D Human Pose Estimation with Bone-based Pose Decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 198–209. [CrossRef]

21. Pose Estimation of a Human Arm Using Kinematic Constraints. Available online: http://www.cvmt.dk/projects/puppet/html/publications/publica-tions.html (accessed on 23 July 2022).
22. Liu, G.Y. Video-Based 3D Human Pose Motion Capture. *J. Comput.-Aided Des. Comput. Graph.* **2006**, *18*, 82–88.
23. Moreno-Noguer, F. 3d human pose estimation from a single image via distance matrix regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2823–2832.
24. Lee, K.; Lee, I.; Lee, S. Propagating LSTM: 3D Pose Estimation Based on Joint Interdependency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 123–141.
25. Wang, J.; Huang, S.; Wang, X.; Tao, D. Not All Parts Are Created Equal: 3D Pose Estimation by Modeling Bi-Directional Dependencies of Body Parts. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7770–7779.
26. Ma, X.; Su, J.; Wang, C.; Ci, H.; Wang, Y. Context Modeling in 3D Human Pose Estimation: A Unified Perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6234–6243.
27. Ashish, V.; Noam, S.; Niki, P. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
28. Angjoo, M.; Michael, J.B.; David, W.J.; Jitendra, M. End-to-end recovery of human shape and pose. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7122–7131.
29. Loper, M. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* **2015**, *34*, 1–16. [CrossRef]
30. Xu, J.; Yu, Z.; Ni, B.; Yang, X.; Zhang, W. Deep Kinematics Analysis for Monocular 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 896–905.
31. Loper, M. An Image Cues Coding Approach for 3D Human Pose Estimation. *ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 1–20.
32. Pavlakos, G.; Zhou, X.; Daniilidis, K. Ordinal Depth Supervision for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7307–7316.
33. Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking. Available online: https://arxiv.org/abs/1904.01324 (accessed on 23 July 2022).
34. Wang, M.; Chen, X.P.; Liu, W.T. DRPose3D: Depth ranking in 3D human pose estimation. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 978–984.
35. Wu, H.; Xiao, B. 3D Human Pose Estimation via Explicit Compositional Depth Maps. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12378–12385.
36. Henry, P. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *Int. J. Robot. Res.* **2013**, *31*, 647–663. [CrossRef]
37. Shotton, J.; Fitzgibbon, A.; Cook, M. Real-time human pose recognition in parts from single depth images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1297–1304.
38. Yi, X.Y. TransPose: Real-time 3D Human Translation and Pose Estimation with Six Inertial Sensors. *ACM Trans. Graph.* **2021**, *40*, 1–13. [CrossRef]