

Survey on Multi-Person 3D Reconstruction from Monocular View

Jingyao Cai¹, Boyuan Cheng¹, Yingjie Xi¹, and Xiaosong Yang¹

National Centre for Computer Animation, Bournemouth University, Poole Dorset
BH12 5BB, United Kingdom
{s5604009, bcheng, yxi, xyang}@bournemouth.ac.uk

Abstract. Reconstructing human pose and shape from a monocular view is a longstanding problem in the field of computer graphics. The goal is to reconstruct the 3D model of human body surface from an image or a video. Most existing monocular human reconstruction methods have achieved great progress which primarily focus on single-person reconstruction. However, research on multi-person reconstruction from a monocular perspective still face certain challenges. In this paper, we comprehensively introduce existing methods in the field of monocular multi-person reconstruction. Additionally, we discuss the challenges and insights into future research directions related to multi-person reconstruction.

Keywords: comprehensive survey · multi-person reconstruction · monocular view.

1 Introduction

Multi-person 3D reconstruction in monocular view aims to generate the 3D full body mesh for all individuals based on monocular view inputs. For conventional methods of making multiple 3D human models involve complex, time-consuming and hand-crafted processes, which present high cost and many obstacles in producing high-quality and fully detailed models. Over the past two decades, considerable research has focused on recovering a single 3D human from single image with significant progress achieved. With these advancements, researchers further pursue to reconstruct multi-human body and shape in 3D space using only monocular input. It is a challenging task for academics in computer graphics. Moreover, it also serves as a core task that benefits corresponding applications such as surveillance, sportscast, gaming and mixed reality.

In this paper, we offer a detailed introduction and a summary of recent advancements in monocular multi-person 3D reconstruction. We firstly delve into a thorough exploration of all existing monocular multi-person reconstruction methods, providing detailed descriptions of algorithmic approaches, datasets, and evaluation metrics. Finally, we discuss the challenges and future directions in the development of multi-person reconstruction field.

2 Multi-Person 3D reconstruction in monocular view

Multi-Person 3D reconstruction in monocular view focus on recovering 3D human mesh from crowded scenes with monocular input. The core lies in the accuracy of mapping the positional relationships of individuals in 2D input to 3D space, as well as the completeness of the models for each individual in the scene. In Table 1, we present a state-of-the-art compilation of existing research papers.

Regarding the classification of methods, more commonly used criteria is based on design strategy which are categorized into two classes: the top-down strategy and the bottom-up strategy.

2.1 The top-down strategy

The top-down strategy initially employs target detection methods to decompose multi-person information in monocular input into individual information. Subsequently, the individual information is fed into the network for the prediction of single-person meshes. Finally, the single person results are combined to get reconstructed multi-person scene. Based on the top-down strategy, methods are capable of achieving relatively comprehensive reconstruction for individual persons. However, they tend to have lower accuracy in handling issues related to the positional information, occlusion and truncation of multiple individuals in the scene.

M3DPSE [9] integrates a feedforward-feedback mechanism involving 2D joint detection, semantic segmentation, and 3D pose prediction. The algorithm refines these predictions using semantic data to align model body parts with image regions, enhancing the accuracy of pose and shape estimations. It incorporates scene constraints by automatically estimating ground plane support and preventing simultaneous volume occupancy. The method also extends to video, optimizing temporal assignments and maintaining coherent motion across frames, which helps in achieving realistic and consistent reconstructions of multiple interacting people.

CRMH [11] employs a coherent method incorporating the SMPL parametric body model, which facilitates handling occlusions and preventing interpenetrations. The algorithm introduces two novel losses: an interpenetration loss to prevent overlapping of reconstructed individuals and a depth ordering-aware loss that promotes accurate depth layering consistent with visual cues. This system operates in a feedforward manner, aiming to produce coherent and non-colliding reconstructions without explicit 3D annotations.

HMAR [12] introduces a novel approach for tracking multiple people in video using 3D representations. It first detects individuals in video frames using bounding boxes, then extracts 3D geometry and appearance information. This data, which includes 3D appearance, pose, and location, is then processed by a transformer that aggregates this information over time, helping to maintain track of each individual across the sequence. This method leverages 3D data to improve tracking accuracy, especially in scenarios with occlusion and viewpoint changes.

Table 1. Overview of multi-person reconstruction in monocular view methods.

Method	Input data	Output model type	Design strategy	Used dataset
M3DPSE [9]	Image&video	Parametric	Top-down	[32],[33]
MubyNet [10]	Image	Parametric	Bottom-up	[32],[33],[34]
CRMH [11]	Image	Parametric	Top-down	[32],[33],[34],[35], [37],[38],[39],[40]
Single-Shot [44]	Image	Parametric	Bottom-up	[32],[34],[35],[40]
Pose2UV [18]	Image&video	Parametric	Top-down	[18],[32],[33],[34],[35], [37],[38],[40],[43]
HMAR [12]	Video	Parametric	Top-down	[32],[34],[35],[39],[46]
REMIPS [13]	Image	Parametric	Top-down	[32],[33],[34],[37],[38], [39],[40],[47],[48]
BSDD [14]	Image	Parametric	Bottom-up	[35],[43]
ROMP [15]	Image	Parametric	Bottom-up	[32],[33],[34],[36],[37], [38],[39],[40],[41],[43]
BMP [16]	Image	Parametric	Bottom-up	[32],[33],[34],[35],[36], [37],[38],[39],[40],[43]
MIR [17]	Image	Clothed human	Bottom-up	[17]
GLAMR [19]	Video	Parametric	Top-down	[19],[43],[49]
OCHMR [20]	Image	Parametric	Top-down	[34],[36],[37],[38], [40],[41],[43]
Multi Person [21]	Image	Parametric	Top-down	[32],[34],[35],[36],[37], [38],[40],[42],[43]
BEV [22]	Image	Parametric	Bottom-up	[22],[32],[33],[34],[37], [38],[40],[41],[42]
3DCrowdNet [23]	Image	Parametric	Top-down	[23],[32],[33],[34], [35],[40],[41],[43]
Human in 4D [24]	Video	Parametric	Bottom-up	[32],[34],[36],[37],[38], [39],[40],[43],[46]
PSVT [25]	Video	Parametric	Bottom-up	[22],[32],[33],[34],[35],[37], [38],[40],[41],[42],[43]
Crowd3D [26]	Image	Parametric	Bottom-up	[26],[32],[35],[42]
Multi-HMR [45]	Image	Parametric	Bottom-up	[33],[35],[42],[43], [45],[50],[51],[52]

REMIPS [13] is a model designed for 3D reconstruction of multiple interacting people from monocular images under weak supervision. It features a novel transformer network that processes person and positional-encoded image feature tokens, enabling the estimation of 3D pose and shape for variable numbers of people. The model innovates with self- and interpenetration-collision models to handle mesh collisions and incorporates self-supervised losses to improve flexibility and generalization in real-world scenarios.

Pose2UV [18] addresses multi-person mesh recovery from a single image by tackling occlusions and extracting valid target features. It utilizes predicted 2D poses to identify individuals and employs a visible pose-mask module to enhance feature extraction. The algorithm introduces a UV map representation, supported by a novel learning-based UV prior, to facilitate human mesh reconstruction. This approach enables handling occlusions and reconstructing plausible human meshes even when only partial body cues are visible.

GLAMR [19] is a method for 3D human mesh recovery from videos taken with dynamic cameras, robust to severe and long-term occlusions. It uses a deep generative motion infiller to reconstruct occluded body motions by leveraging visible movements and a global trajectory predictor to establish consistent global coordinates. This method also integrates a global optimization framework that aligns camera poses and human trajectories with video evidence, significantly improving accuracy in handling occlusions and camera movement.

OCHMR [20] is designed to address challenges in multi-person 3D mesh recovery from images with significant occlusion. It integrates spatial-context in the form of body-center maps into the recovery process, utilizing CoNorm blocks within its architecture to adaptively modulate feature maps based on this context. This method not only improves the disambiguation of overlapping humans but also enhances the robustness and accuracy of mesh predictions in crowded scenes.

MultiPerson [21] uses a coarse-to-fine pipeline to reconstruct multi-person 3D meshes from monocular image. It starts by estimating robust 3D skeletons for each person, then applies inverse kinematics to transform these skeletons into deformable mesh parameters. Finally, a relation-aware Transformer refines these parameters by considering intra- and inter-personal relationships, enhancing the accuracy and realism of the reconstructed meshes.

3DCrowdNet [23] leverages a 2D pose estimator to obtain the 2D key-point positions of each individual and avoid the domain gap typically present with motion capture datasets. It utilizes a joint-based regressor to sample image features from predicted joint locations to preserve spatial activation and effectively isolate target features from others in the scene. This approach ensures that 3DCrowdNet accurately distinguishes a target individual’s features even amidst significant inter-person occlusion.

2.2 The bottom-up strategy

The Bottom-up Strategy takes a holistic approach to handle multi-person information and achieve multi-person reconstruction. Method under this strategy

feeds the entire monocular multi-person input into the network for simultaneous processing, allowing for a more comprehensive understanding of corresponding information such as positional occlusion between multiple individuals. However, due to the input being processed in a holistic manner rather than individual character information, the effectiveness in generating the detailed completeness of individual characters is significantly reduced.

MubyNet [10] is a multitask framework for simultaneous localization, 3D pose and shape estimation of multiple people from monocular images. It identifies body joints and limbs, grouping them into person instances using a learned scoring function that combines 2D and 3D information. The core feature is the integration of 3D pose prediction in a single model, solving a binary integer linear program for person grouping under kinematic constraints. This method operates without prior knowledge of the number of people or their visibility relations in the scene.

Single-Shot [44] is an end-to-end model that performs real-time 3D reconstruction of multiple persons from a single image. Each grid cell in the output tensor contains personal details, enabling the prediction of both 3D shapes and root joint depths without requiring bounding box input. It efficiently handles multiple individuals simultaneously, maintaining constant complexity regardless of the number of persons in the image.

BSDD [14] presents an algorithm to address the inherent body size and depth ambiguity in multi-person reconstruction from a single image. It uniquely enforces that individuals' feet remain on the ground, thereby disambiguating body scale and depth. The method uses a novel constraint for feet-to-ground distance and optimizes body scale and camera translation relative to the ground plane, significantly improving spatial arrangement and scale accuracy across images.

ROMP [15] aims to regress all 3D body meshes for multiple individuals in a given image in a single stage. It provides an efficient and occlusion-resistant per-pixel representation by predicting Body Center heatmaps and Mesh Parameter maps. By focusing on the body centers and leveraging the pixel-level detail, ROMP enhances robustness and efficiency in extracting 3D meshes even in crowded scenes.

BMP [16] is a single-stage model for multi-person 3D body mesh estimation. It represents each person as a point in spatial-depth space, associating with a body mesh, allowing concurrent localization and mesh estimation. It designs an ordinal depth loss for depth-coherent mesh estimation and a keypoint-aware occlusion augmentation to handle occlusions effectively, significantly enhancing both efficiency and performance in multi-person body mesh estimation.

MIR [17] is a model-free algorithm that achieves spatially coherent reconstructions of multiple clothed individuals from a single image. This end-to-end learning framework leverages multitask networks to estimate both the 3D geometry and 6DOF spatial locations of individuals, handling complex poses, occlusions, and diverse clothing without manual intervention. The method integrates

Table 2. Chronological overview of main datasets utilized by existing monocular multi-person reconstruction algorithms, along with their sizes.

Dataset	Size
LSP [37] & LSPEExtended [38]	Contains 2K images of sportspersons gathered from Flickr and each annotated with 14 joint locations
Human3.6M [32]	Contains 3.6 million human poses and corresponding images
COCO [34]	Contains 2.5 million labeled instances in 328K images
MPII [40]	Contains 25K images containing over 40K people, 410 different human activities
Panoptic [33]	Contains 65 sequences (5.5 hours) and 1.5 millions of 3D skeletons
MPI-INF-3DHP [36]	Contains over 1.3M frames captured from the 14 cameras including 8 actors performing 8 activities
MuPoTS-3D [35]	Contains more than 8K frames from 20 real-world scenes with up to three subjects
PoseTrack [39]	Contains 514 videos over 150,000 annotated poses including 66,374 frames in total
AVA [46]	Contains annotates 80 atomic visual actions in 430 15-minute movie clips, resulting in 1.62M action labels
3DPW [43]	Contains 60 video sequences and 18 3D models in different clothing variations
MuCo3DHP [35]	Contains approximately 400K composite frames
Crowdpose [41]	Contains about 20K images and a total of 80K human poses with 14 labeled keypoints
AMASS [49]	Contains more than 40 hours of motion data, spanning over 300 subjects, more than 11K motions
EHF [50]	Contains 100 frames of one subject, showcasing diverse body poses with natural finger and facial articulation
FlickrCI3D [48]	Contains 55,095 images of 90,167 pairs of people in interaction scenarios
CHI3D [48]	Contains 631 sequences containing 2,525 contact events 728,664 ground truth 3d poses
FlickrSC3D [47]	Contains 18,187 images of 24,312 people, classified by annotators in 3 self-contact classes.
MPSD [17]	Contains 450k pairs of images and 3D models, each scene rendered into 16 camera views at 512×512 resolution
AGORA [42]	Contains 173K individual person crops and provides SMPL/SMPL-X parameters and segmentation masks
3DMPB [18]	Contains more than 10K images and accurate 3D annotations
RH [22]	Contains 7.6K images with weak annotations of over 24.8K people
3DPWCrowd [23]	Contains 1073 images and 1923 persons with GT 3D pose and shape annotations
LargeCrowd [26]	Contains over 100K labeled crowded people in 733 gigapixel large-scene images (19200×6480)
UBody [51]	Contains 1000K images with 2D keypoints annotation and 3D SMPLX annotation
BEDLAM[52]	Contains approximately 380K unique image frames with 1 to 10 people each, totaling 1 million bounding boxes
CUFFS[45]	Contains 60k images featuring synthetic renderings of people with close-up views of full bodies and clear hands

depth and instance segmentation to enhance the reconstruction accuracy and spatial coherence of each individual in the scene.

BEV [22] introduces an imaginary 2D bird’s-eye-view alongside a traditional front-view to enable depth reasoning. It uses body center heatmaps and localization offset maps from these two views, and combines them to generate 3D center/offset maps for final mesh regression. This approach allows for effective disambiguation of depth and improves localization of multiple people, even in the presence of severe occlusions.

Humans in 4D [24] utilizes a fully transformer-based architecture, HMR 2.0, for recovering 3D human meshes from single images and tracking them over time in videos. By utilizing the 3D reconstruction results from HMR 2.0, it constructs the 4DHumans system which can track multiple individuals in videos and maintain continuity of identities through occlusion events.

PSVT [25] is an end-to-end framework using Progressive Video Transformers. It utilizes a spatio-temporal encoder to understand global feature interactions and employs progressive pose and shape decoders to handle pose estimation and mesh reconstruction simultaneously. The key innovation is the use of progressive decoding, which updates the pose and shape queries frame by frame, and pose-guided attention to enhance the accuracy of shape estimations.

Crowd3D [26] is a framework for reconstructing 3D poses, shapes, and locations of hundreds of people from a single image. It introduces a Human-scene Virtual Interaction Point (HVIP) to transform complex crowd localization into pixel localization, aiding in global consistency. The approach includes an adaptive human-centric cropping scheme to manage varying human scales and a progressive reconstruction network to integrate scene-level camera and ground plane predictions, ensuring spatial coherence and addressing depth ambiguity effectively.

Multi-HMR [45] employs a Vision Transformer (ViT) backbone to detect multiple humans and regress their whole-body pose, shape, and spatial location. This approach uniquely integrates a prediction head using cross-attention, called the Human Prediction Head (HPH), to enhance pose and shape parameter prediction. Multi-HMR also optionally adjusts for camera intrinsics, enhancing placement accuracy in camera space.

3 Dataset

Datasets used in single-view multi-person reconstruction typically include a set of monocular images or videos containing multiple individuals 3D mesh, each with corresponding 3D pose or shape information. Due to the close relationship between monocular multi-person reconstruction algorithms and single-person reconstruction, pose estimation and tracking, as well as human segmentation, the types of datasets used are numerous and diverse. Table 2 summarizes most of the datasets currently used in monocular multi-person reconstruction research, detailing the main tasks and scale of each dataset. Additionally, Table 1 also shows the specific datasets relied upon by various methods.

Table 3. Overview of main evaluation metrics utilized by existing multi-person reconstruction algorithms.

Metrics	Detail&Measurement
MPJPE	Mean Per Joint Position Error: To evaluate the inferred pose centered in its hip joint.
PA-MPJPE	Procrustes-aligned MPJPE: To measure the accuracy of 3D human pose estimation by calculating MPJPE after using Procrustes Analysis to perform 3D alignment.
G-MPJPE	Global-MPJPE: Calculate MPJPE by placing the SMPL model in global coordinates.
PVE	Per-Vertex Error: To evaluate the 3D surface error by measuring the Euclidean distance between the predicted vertices and the ground truth.
PA-PVE	Procrustes-aligned PVE: Measure PVE after Procrustes Analysis to perform 3D alignment.
G-PVE	Global-PVE: To calculate per-vertex error in the global coordinates.
MPVPE	Mean Per Vertex Position Error: To evaluate 3D reconstruction or pose estimation accuracy, quantifies the average distance between predicted and actual 3D vertices.
MVE	Mean Vertex Error: Calculate Euclidean distance between ground truth and estimated mesh vertices.
NMVE	Normalized Mean Vertex Error: Normalized MVE by F1 score to punish misses and false alarms in the detection.
NMJE	Normalized Mean Joint Error: Normalized MPJPE by F1 score to punish misses and false alarms in the detection.
MPJAE	Mean Per Joint Angle Error: To measure the angle in degrees between the predicted part orientation and the ground truth orientation.
PA-MPJAE	Procrustes-aligned MPJAE: To measure MPJAE after rotating all predicted orientations by the rotation matrix obtained from the procrustes matching step.
PCK	Percentage of Correct Keypoints: To measure if the predicted keypoint and the true joint are within a certain distance threshold.
3DPCK	3D Percentage of Correct Keypoints: To measure keypoint accuracy by considering if their Euclidean distance from the true keypoint is below a specified threshold.
AUC	Area Under the PCK-threshold: To evaluate the mean accuracy within error range.
MOTA	Multi-Object Tracking Accuracy: To evaluate the overall performance of multi-object tracking by summarizing errors from false positives, false negatives and ID switches.
HOTA	Higher Order Tracking Accuracy: To evaluate multi-object tracking outcome by calculating the alignment between predicted and actual trajectories over multiple frames.
DOAL	Depth Ordering-aware Loss: To quantify the percentage of correctly estimated ordinal depth relations between all pairs of people in the image.
DHC	Discrete height comparison: Represent the percentage of correctly estimated ordinal height relations between all pairs of people in the image.
P2S	Point To Surface error: To measure the average distance from points on one surface to another to assess surface reconstructed accuracy.
2D/3D IoU	Intersection of Union: To evaluate reconstruction accuracy by assessing the overlap between the predicted volume and the ground truth volume.
PPDS	Pair-wise Percentual Distance Similarity: To evaluate the spatial distribution of crowd, use the torso center as the location of people and calculate distances.
PA-PPDS	Procrustes-Aligned Pair-wise Percentual Distance Similarity: Align reconstructed crowd by Procrustes alignment to exclude influence of scale and rotation.
MRPE	Mean Root Position Error: Measures the Euclidean distance between the estimated and ground truth root joints in the camera coordinate system.
PCOD	Percentage of Correct Ordinal Depth: To evaluate the ordinal depth relations between all pairs of people in the image.

4 Evaluation

Evaluation metrics are utilized to quantify the performance of models and the quality of their predictions. The selection of these metrics typically depends on the characteristics of the dataset, the type of task, and the specific objectives of the research. For the evaluation of monocular multi-person reconstruction methods, we present the key metrics used to assess reconstruction results in Table 3. It provides detailed explanations of the assessment criteria for each metric along with the corresponding methods that employ these metrics for evaluation.

5 Conclusion

In this survey, we provide a detailed overview of all current multi-person reconstruction techniques and discuss the associated datasets and evaluation criteria.

Limitation: We observe that multi-person reconstruction from a single viewpoint is a relatively emerging research area, but current methods still face limitations. When estimating multiple people from a single viewpoint, the information provided by a single photo or video is limited. Accuracy in predicting the spatial relationship from 2D views to 3D spaces often contains errors, especially in complex or crowded environments. Obstructions—whether from the individuals themselves, between people, or between people and their environment—can compromise the accuracy and robustness of reconstructions. Moreover, most multi-person reconstruction methods are based on parametric model techniques which is primarily effective for figures wearing minimal clothing. It tends to be less precise for individuals with exaggerated body shapes or loose clothing. Methods that directly reconstruct human figures with detailed features like clothing, hair, and expressions are easily influenced by environmental factors, leading to poor generalization.

Discussion: Therefore, in the future of multi-person 3D reconstruction, developing more flexible representation methods should become a primary focus. First and foremost, accurately learning and understanding the spatial relationships among multiple individuals, as well as more effectively addressing the challenges posed by occlusions, are critical priorities for multi-person reconstruction techniques. Secondly, to enhance the accuracy and generalizability of this work, creating more diverse and suitable datasets is also essential. Additionally, further research should be dedicated to exploring how to maintain the generalization features provided by parametric models while effectively reconstructing more realistic human models that include details such as clothing, hair, and facial features. This will not only improve the practicality of the models but also advance multi-person 3D reconstruction technology to higher levels of application.

References

1. Saito S, Huang Z, Natsume R, et al. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2304-2314.
2. Xiu Y, Yang J, Tzionas D, et al. Icon: Implicit clothed humans obtained from normals[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 13286-13296.
3. Ju S X, Black M J, Yacoob Y. Cardboard people: A parameterized model of articulated image motion[C]//Proceedings of the Second International Conference on Automatic Face and Gesture Recognition. IEEE, 1996: 38-44.
4. Marr D, Nishihara H K. Representation and recognition of the spatial organization of three-dimensional shapes[J]. Proceedings of the Royal Society of London. Series B. Biological Sciences, 1978, 200(1140): 269-294.
5. Wang M, Qiu F, Liu W, et al. Monocular human pose and shape reconstruction using part differentiable rendering[C]//Computer Graphics Forum. 2020, 39(7): 351-362.
6. Sminchisescu C, Triggs B. Estimating articulated human motion with covariance scaled sampling[J]. The International Journal of Robotics Research, 2003, 22(6): 371-391.
7. Plänkers R, Fua P. Tracking and modeling people in video sequences[J]. Computer Vision and Image Understanding, 2001, 81(3): 285-302.
8. Kakadiaris L, Metaxas D. Model-based estimation of 3D human motion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1453-1459.
9. Zanfır A, Marinoiu E, Sminchisescu C. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2148-2157.
10. Zanfır A, Marinoiu E, Zanfır M, et al. Deep network for the integrated 3d sensing of multiple people in natural images[J]. Advances in neural information processing systems, 2018, 31.
11. Jiang W, Kolotouros N, Pavlakos G, et al. Coherent reconstruction of multiple humans from a single image[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5579-5588.
12. Rajasegaran J, Pavlakos G, Kanazawa A, et al. Tracking people with 3D representations[J]. arXiv preprint arXiv:2111.07868, 2021.
13. Fieraru M, Zanfır M, Szente T, et al. REMIPS: Physically consistent 3D reconstruction of multiple interacting people under weak supervision[J]. Advances in Neural Information Processing Systems, 2021, 34: 19385-19397.
14. Ugrinovic N, Ruiz A, Agudo A, et al. Body size and depth disambiguation in multi-person reconstruction from single images[C]//2021 International Conference on 3D Vision (3DV). IEEE, 2021: 53-63.
15. Sun Y, Bao Q, Liu W, et al. Monocular, one-stage, regression of multiple 3d people[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 11179-11188.
16. Zhang J, Yu D, et al. Body meshes as points[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 546-556.
17. Mustafa A, Caliskan A, Agapito L, et al. Multi-person implicit reconstruction from a single image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14474-14483.

18. Huang B, Zhang T, Wang Y. Pose2uv: Single-shot multiperson mesh recovery with deep uv prior[J]. *IEEE Transactions on Image Processing*, 2022, 31: 4679-4692.
19. Yuan Y, Iqbal U, Molchanov P, et al. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022: 11038-11049.
20. Khirodkar R, Tripathi S, Kitani K. Occluded human mesh recovery[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022: 1715-1725.
21. Cha J, Saqlain M, Kim G U, et al. Multi-person 3d pose and shape estimation via inverse kinematics and refinement[C]//*European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022: 660-677.
22. Sun Y, Liu W, Bao Q, et al. Putting people in their place: Monocular regression of 3d people in depth[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 13243-13252.
23. Choi H, Moon G, Park J K, et al. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 1475-1484.
24. Goel S, Pavlakos G, Rajasegaran J, et al. Humans in 4d: Reconstructing and tracking humans with transformers[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023: 14783-14794.
25. Qiu Z, Yang Q, Wang J, et al. PSVT: End-to-end multi-person 3D pose and shape estimation with progressive video transformers[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023: 21254-21263.
26. Wen H, Huang J, Cui H, et al. Crowd3D: Towards hundreds of people reconstruction from a single image[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023: 8937-8946.
27. Li X, Huang J, et al. Learning to infer inner-body under clothing from monocular video[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2022.
28. Doersch C, Zisserman A. Sim2real transfer learning for 3d human pose estimation: motion to the rescue[J]. *Advances in Neural Information Processing Systems*, 2019, 32.
29. Wei W L, Lin J C, Liu T L, et al. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 13211-13220.
30. Lee G H, Lee S W. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021: 12375-12384.
31. Peng B, Hu J, Zhou J, et al. Selfnerf: Fast training nerf for human from monocular self-rotating video[J]. *arXiv preprint arXiv:2210.01651*, 2022.
32. Ionescu C, Papava D, Olaru V, et al. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2013, 36(7): 1325-1339.
33. Joo H, Liu H, Tan L, et al. Panoptic studio: A massively multiview system for social motion capture[C]//*Proceedings of the IEEE international conference on computer vision*. 2015: 3334-3342.
34. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//*Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer International Publishing, 2014: 740-755.

35. Mehta D, Sotnychenko O, Mueller F, et al. Single-shot multi-person 3d pose estimation from monocular rgb[C]//2018 International Conference on 3D Vision (3DV). IEEE, 2018: 120-130.
36. Mehta D, Rhodin H, Casas D, et al. Monocular 3d human pose estimation in the wild using improved cnn supervision[C]//2017 international conference on 3D vision (3DV). IEEE, 2017: 506-516.
37. Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation[C]//bmvc. 2010, 2(4): 5.
38. Johnson S, Everingham M. Learning effective human pose estimation from inaccurate annotation[C]//CVPR 2011. IEEE, 2011: 1465-1472.
39. Andriluka M, Iqbal U, Insafutdinov E, et al. PoseTrack: A benchmark for human pose estimation and tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5167-5176.
40. Andriluka M, Pishchulin L, Gehler P, et al. 2d human pose estimation: New benchmark and state of the art analysis[C]//Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014: 3686-3693.
41. Li J, Wang C, Zhu H, et al. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10863-10872.
42. Patel P, Huang C H P, Tesch J, et al. AGORA: Avatars in geography optimized for regression analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13468-13478.
43. Von Marcard T, Henschel R, Black M J, et al. Recovering accurate 3d human pose in the wild using imus and a moving camera[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 601-617.
44. Kim S H, Chang J Y. Single-shot 3d multi-person shape reconstruction from a single rgb image[J]. Entropy, 2020, 22(8): 806.
45. Baradel F, Armando M, Galaaoui S, et al. Multi-HMR: Multi-Person Whole-Body Human Mesh Recovery in a Single Shot[J]. arXiv preprint arXiv:2402.14654, 2024.
46. Gu C, Sun C, Ross D A, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6047-6056.
47. Fieraru M, Zanfir M, Oneata E, et al. Learning complex 3D human self-contact[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(2): 1343-1351.
48. Fieraru M, Zanfir M, Oneata E, et al. Three-dimensional reconstruction of human interactions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7214-7223.
49. Mahmood N, Ghorbani N, Troje N F, et al. AMASS: Archive of motion capture as surface shapes[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 5442-5451.
50. Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3d hands, face, and body from a single image[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10975-10985.
51. Lin J, Zeng A, Wang H, et al. One-stage 3d whole-body mesh recovery with component aware transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 21159-21168.
52. Black M J, Patel P, Tesch J, et al. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 8726-8737.