Constrained Humanification: Improving Multi-Person Reconstruction Using Temporal Constraints

Shefali Srivastava
Carnegie Mellon University
shefalis@andrew.cmu.edu

Anirudh Chakravarthy
Carnegie Mellon University
achakrav@andrew.cmu.edu

Figure 1. We demonstrate the problem with single-frame based approaches. The highlighted individual is a challenging case for existing multi-person reconstruction methods since the person is heavily occluded. This can be alleviated by using optical flow to incorporate the motion over time.

1. Introduction

Recently, vision tasks in 3D have achieved monumental performances for tasks such as 3D keypoint estimation [12, 28], 3D pose and shape estimation [6, 9, 10], 3D reconstruction [32] and full-body shape recovery. 3D reconstruction in isolation is much exploited in literature but the real strength of the task comes via scene estimation and reconstruction that aids the process of 3D human pose estimation and reconstruction.

For multi-person pose estimation, there are usually two approaches considered in literature. Bottom-up approaches first detect 2D key-points or body joints in the scene and then use them as handles to combine for a coherent 3D reconstruction suggestive of one human. Bottom up approaches are complicated for other 3D representations such as mesh recoveries and parametric representations. Top-down approaches rely on 2D object detectors to first detect objects and individual persons in a scene and then perform 3D pose estimation on each person separately. These top-down approaches are able to collectively harness the superior performances of state of the art 2D object detectors and as reported in [12], leverage smooth pipelines for 2D to 3D key-point regressions. While these perform well, there is still scope for improvement, specially in cases of reconstructions with occlusions. Inconsistent depth ordering is another very common issue along with inter-penetrations. Due to these shortcomings, it is critical to go beyond just predicting a reasonable 3D pose for each person individually, and instead estimate a coherent reconstruction of all the persons in a scene.

Recent works have emerged which perform multi-person reconstruction [8]. In such approaches, a top-down approach is employed where in the first stage, 2D Object detection is performed to detect each person, and then the 3D human meshes are regressed for each detected person. In addition, several other constraints have been explored to enforce coherent reconstruction. For example, interpenetration loss penalises reconstructions whose meshes intersect with each other. Depth-ordering aware losses have also been used to resolve ambiguities.

However, to the best of our knowledge, no work exists to disambiguate occlusion. One important cue which is often overlooked in context of 3D pose estimation is that of temporal information. Temporal information can be very useful to reconstruct multiple occluded instances of people.

Consider the scenario in Fig 1. The highlighted individual moves from being almost entirely occluded to being partially occluded. Performing reconstruction in the both the frames individually can be very challenging. However, if we are able to propagate the optical flow from one frame to another, the motion of the person over time can help to better reconstruct the individual.

In this work, we experiment with leveraging optical flow as a cue for 3D pose estimation. We propose several fusion methods to augment the per-frame detections with temporal information. We also demonstrate, both quantitatively and qualitatively, the promise of using such an approach.

2. Related Work

2.1. 3D Pose Estimation

In present day literature, various representational forms of 3D have been exploited for Human Pose Estimation. Learning based method estimate 3D either from detected 2D key points or images. Some works tackle the problem using multiple image inputs [18,23] while some recent ones try to estimate a human pose using single RGB image augmented with depth information [16, 25, 26]. While estimation of 3D pose for single human is challenging in itself, multi-human pose estimation is also a much explored task in recent literature due to its relevance to the real world scenario. This work specially targets multi-human pose estimation by coherently considering the whole scene. It builds on the works of [8] where Mask R-CNN is used to extract 2D bounding boxes for every person in the image. Meshes are used to represent the 3D reconstructions. For mesh representation, SMPL parameters are regressed. They refer to it as SMPL R-CNN and apply two novel losses referred to as the interpenetration loss and the depth-ordering consistency loss. The interpenetration loss penalises the 3D reconstruction where a part of the reconstruction occupies an area already occupied by another reconstruction. The depth-ordering consistency loss projects the 3D reconstruction onto an image place and uses mask segmentation available for the image to penalise any reconstruction whose projection does not match its segmented image. This way, depth consistency is maintained and learnt by the network. We aim to build on this work by augmenting this network with optical flow information to handle possible future occlusions.

2.1.1 Single-person Pose Estimation

Multiple representations for 3D pose estimation have gained recognition throughout literature. Some recent works tackle the problem by estimating 3D poses in the form of skeletons [12, 15, 20, 22, 28, 30, 31, 34]. Some also try to do it in a non-parametric way by exploiting 3D shapes [4,27,32]. SMPL [11] has been adopted as a state-of-the-art method off-late, which estimates 3D poses in the form of a human mesh. SMPL is a skinned, vertex-dependent model that is capable of representing various human poses. The parameters of the mesh are learned from data. The model learns to estimate the shape β, pose θ, blend weights, and a regressor from vertices to 2D joint locations. Many automatic approaches such as SMPLify [3] and SMPL-X [19] has been built on top of SMPL to iteratively fit 2D joints to a 3D mesh model for a human. These models are much more expressive than SMPL. There are also other models that use silhouettes [21] and voxel occupancy grids [32] but are not as popular.

Although these methods have been somewhat successful, learning based methods that rely on only input images have also gained traction due to successful deep learning based visual feature extractors. These networks help 3D mesh estimation by directly regressing the shape and pose from images. By letting the network train end-to-end, hand-designed features that can potentially act as bottlenecks are avoided. Visual feature extractors via standard backbone networks also helps leverage pre-trained models on large datasets. The extracted features can be in any format. Multiple format have again been exploited in literature to act as intermediates for this task.

Keypoints and Silhouettes are used in [21], semantic part segmentation is used in [17]. [10] regress the mesh vertices to Graph Convolutional Neural Networks. Using a prior over the 3D shapes during training can penalize imperfect reconstructions. This idea is used by [9].

We utilise an idea similar to [2] and use temporal context to improve the regression network.

2.1.2 Multi-person Pose Estimation

Similar to how a single-person pose-detection is performed, a multi-person pose detection sometimes involves isolation of the single-person first and then estimating 3D using single-person pose estimation techniques. This top-down approach is usually quite popular with 3D pose estimation methods and different techniques and architectures are adopted in literature to extract single-persons from an image and estimate their respective 3D representations. Many works build on frameworks such as R-CNN and Mask R-CNN [5,7] to segment out persons. LCR-Net [24] regresses joint offsets to classify poses. [33] uses an interesting approach to incorporate a scene prior for iteratively improve the 3D shapes and poses of people. We seek inspiration from this work and try to incorporate scene constraints along with a novel proposal of optical flow. Our approach is instead a top-down approach where SMPL parameters are regressed using the current image frame and the optical flow to the current image frame. We believe that information

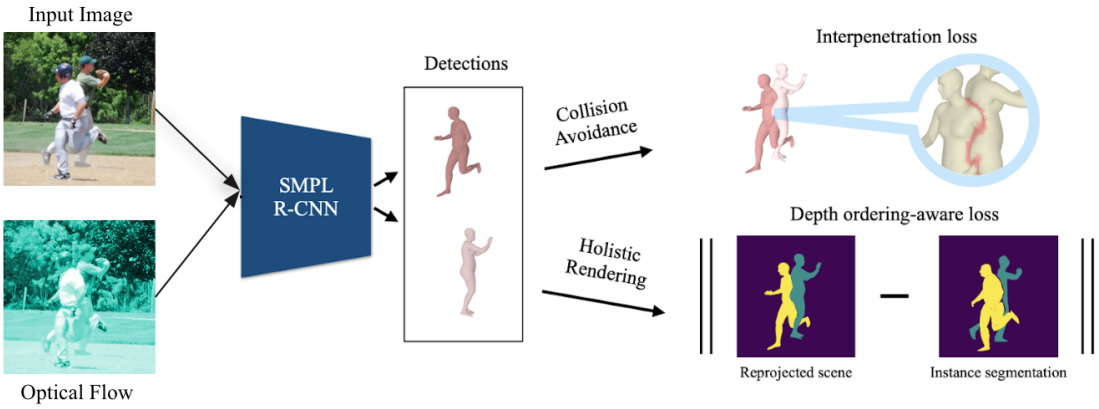


Figure 2. Our proposed network architecture. In addition to image inputs, SMPL-RCNN [8] is also given optical flow as input to improve multi-person reconstruction.

contained in the optical flow of a frame helps the network understand the 3D shape better and build a better prior in the network for occluded objects. Even if an object is not visible in the next frame which can potentially hinder the 3D reconstruction of it, we believe the information from the optical flow can somehow be helpful in preserving knowledge about the movement of the person. This can help 3D reconstruction of the human from the next frame.

2.2. Optical Flow

Optical flow is the task of estimating per-pixel motion between video frames. It has been a long-standing task riddles with problems involving fast-movements, motion blurs and occlusions. As is with most vision tasks, estimation of optical flow has slowly moved from being computed via either hand-crafted features or complicated mathematical equations to being learning based. In this work, use Recurrent All-Pairs Field Transforms (RAFT) [29] to estimate optical flow from video frames. RAFT utilises a feature encoder to extract feature vectors for every pixel in the image. A correlation layer on top of it worked to produce a 4 dimensional volume for all pairs of pixels. Pooling layers help reduce dimensions. The core strength of the algorithm lies in Recurrent Neural Network based update operator that iteratively updates the flow field. It has been shown to have state of the art accuracy, high efficiency and strong generalisation capabilities despite being trained only on synthetic data.

3. Proposed Method

For a given image $\mathbf{I}$, multi-person reconstruction aims to predict a set of SMPL parameters [11] $\{(\theta_1, \beta_1), \cdots (\theta_i, \beta_i) \cdots (\theta_K, \beta_K)\}$ for each person $i \in [1, K]$ in $\mathbf{I}$, where K is the number of people in the scene. In this section, we propose an optical flow-based supervision mechanism to improve the performance of existing multi-person reconstruction networks.

3.1. Optical Flow Generation

In order to augment the network with temporal information, we need to generate optical flow. However, existing datasets do not contain optical flow supervision. To this end, we use RAFT [29], a state-of-the-art optical flow network.

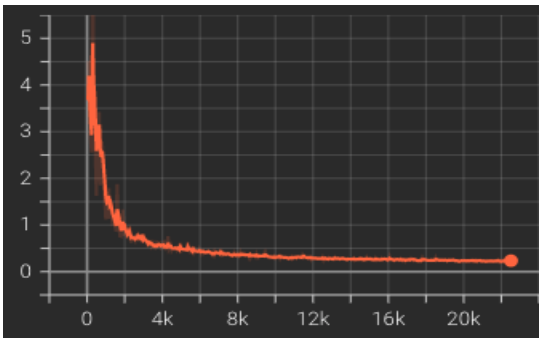
Given two frames $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$, RAFT estimates a dense displacement field $\mathbf{F}_{t-1\to t} = (\mathbf{f}^1, \mathbf{f}^2)$ for each pixel in $\mathbf{I}_{t-1}$. Concretely, each pixel coordinate (u, v) in $\mathbf{I}_{t-1}$ is mapped to the coordinates $(u', v') = (u + f^1(u), v + f^2(v))$ in $\mathbf{I}_t$. We use the generated flow as an input to our network.

3.2. Flow Fusion

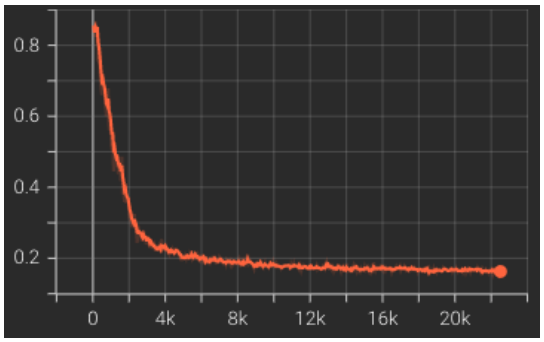
Our next aim is to use the generated flow $\mathbf{F}_{t-1\to t}$ as input to the reconstruction network in addition to the image $\mathbf{I_t}$. Our proposed network architecture is illustrated in Fig 2. In this work, we only focus on early fusion methods.

In order to incorporate optical flow into the single-frame reconstruction network, we proposed two schemes, based on (i) type of input, and (ii) method of fusion.

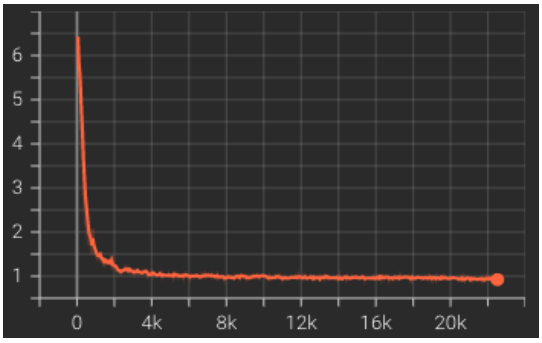
Type of Input. Our first scheme to augment optical flow into the existing network is based on the type of input to the network. Under this scheme, we propose two methods for utilizing optical flow. Our first method naively feeds the generated optical flow $\mathbf{F}_{t-1\to t}$ in addition to the image $\mathbf{I}_t$ as inputs to the network. Alternatively, under the second method, we use the image $\mathbf{I}_{t-1}$ and optical flow $\mathbf{F}_{t-1\to t}$ to generate a warped image $\hat{\mathbf{I}}_t$. We feed this warped image $\hat{\mathbf{I}}_t$



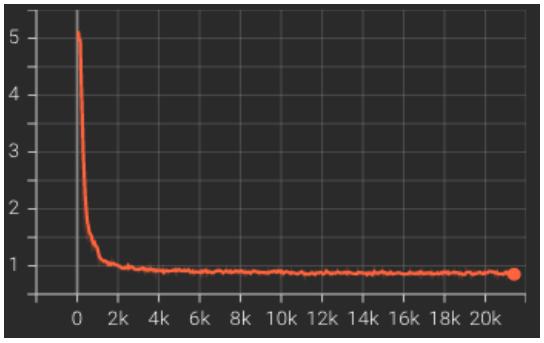
(a) Add Flow



(b) Add Warped Image



(c) Concatenate Flow



(d) Concatenate Warped Image.

Figure 3. Training Loss Curves for 5 epochs.

in addition to $\mathbf{I}_t$ as inputs to our network. This helps the network understand how the objects have moved from the previous frame to the current, which may help disambiguate occlusion.

Method of Fusion. Under the second scheme, we propose two methods to fuse the image and flow-augmented input that can either be flow or the warped image. First, given the flow-augmented input ($\mathbf{F}_{t-1\to t}$ or $\hat{\mathbf{I}}_t$), we use a convolutional layer similar to the first convolutional layer in ResNet-50 to extract features from this input. The number of input channels to the convolutional layer is determined by the type of input (2 channel for optical flow and 3 channel for warped image). Once we extract features from the flow-augmented input, we similarly extract image features using the first convolutional layer of ResNet-50. Using the flow-augmented features and the image features, we either add or concatenate these features before feeding them into the subsequent feature extraction layers in ResNet-50. In order to perform concatenation, we introduce another convolution layer to reduce the number of channels before feeding them into the subsequent layers of the backbone network. The rest of the network is the same as SMPL R-CNN [8].

4. Experiments

In this section, we evaluate the performance of our proposed method.

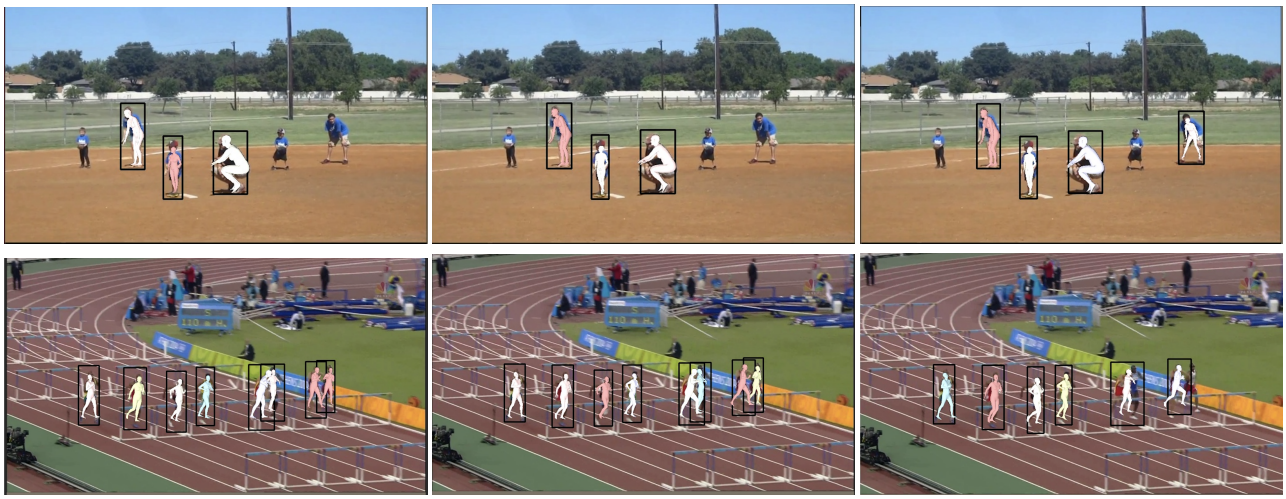
Method	All ↑	Matched ↑	Collisions ↓
SMPL R-CNN [8]	85.32	**89.09**	62
Add Flow	**85.60**	89.01	89
Add Warp	85.27	88.72	59
Concat Flow	53.68	73.92	12
Concat Warp	64.72	78.32	**7**

Table 1. **Results on MuPoTS-3D**. We report the overall 3DPCK accuracy (All), the 3DPCK accuracy only for person annotations matched to a prediction (Matched), and number of collisions.

4.1. Datasets

To train our network, we use the PoseTrack Dataset [1]. It consists of multiple frames corresponding to each sequence with several people in each frame. 2D pose annotations are available and we only use this 2D supervision to train our network.

To evaluate our network, we use the MuPoTS-3D dataset [14]. It also consists of multiple people across multiple frames in a video sequence. It contains 3D pose annotations which makes evaluation of our network feasible on this benchmark.



SMPL R-CNN [8] Add Flow Add Warped Image

Figure 4. We compare the pose estimation performance of the baseline with flow addition and warped image addition.

SMPL R-CNN [8] Add Flow Add Warped Image

Figure 5. We compare the depth ordering of the baseline with flow addition and warped image addition for the corresponding images in Fig 4.

4.2. Implementation Details

The input images to RAFT are resized to 520×960 to compute the optical flow at this resolution. Then, the optical flow and corresponding images are resized to the desired shape for training. We train our network using a batch size of 4 on a single NVIDIA RTX 2080 GPU. We train the addition scheme networks for 5 epochs, and the concatenation scheme networks for 20 epochs.

Since training the network from scratch is not feasible, we used the pre-trained checkpoints from SMPL R-CNN. The entire network was frozen except for the first two back-

bone convolutional layers and the final SMPL regression layer, which were finetuned. We follow the same hyperparameter configuration as SMPL R-CNN.

For evaluation, we use an unofficial Python reimplementation[1] on MuPoTS-3D. Therefore, for fair comparison, we recompute the baseline results reported in [8] using this implementation.

4.3. Quantitative Results

We perform quantitative analysis of our proposed network on the MuPoTS-3D benchmark in Table 1. We use the 3D extension for Percentage of Correct Keypoints [13] (3DPCK) with a threshold of 150 mm to evaluate our network.

On incorporating temporal augmentation by adding flow into the image features, we observe a slight improvement over the baseline performance on the overall 3DPCK accuracy. However, the number of collisions is much higher and the matched 3DPCK accuracy is about the same. This suggests that adding optical flow is useful to regress keypoint locations, but leads to more incoherent or overlapping reconstructions (especially at the non-joint locations).

On adding the warped image, we observe fewer collisions with similar 3DPCK accuracy, which suggests that incorporating explicit information of motion of objects in the previous frame indeed helps disambiguate depth ordering among occluded people.

However, on performing concatenation, we observe significantly lower 3DPCK accuracy. On observing the training loss curves in Fig 3, we hypothesize that the network has not been trained sufficiently. SMPL R-CNN [8] was trained through an extensive pre-training strategy for over 100 epochs. On the other hand, our new flow and fusion branches are trained from scratch for around 20 epochs, which leads to significantly poor performance.

The number of collisions is much lower, which suggests that the network has not learnt to detect humans and regress pose parameters well enough to encounter cases of depth ordering inconsistencies. However, overcoming this challenge requires significant compute and remains beyond the scope of this work.

Interestingly, we observe that running evaluation using the pre-trained checkpoint leads to much higher collisions on MuPoTS-3D than reported in the paper. The difference in 3DPCK is explained due to the implementation differences between the original evaluation script (in Matlab) versus the unofficial re-implementation (in Python).

4.4. Qualitative Results

In this subsection, we examine the qualitative performance of our method with respect to the baseline network in Fig 4 and Fig 5. Since the concatenation scheme was significantly worse, we only visualize the networks following the addition scheme.

As seen in the first row, using warped image features leads to the detection of the player in the top-right. This indicates the success of our approach since we are able to detect additional people using temporal information who were previously not detected. Similarly, in the second row, the runners at the right end are not reconstructed well by SMPL R-CNN. Using the optical flow improves the relative depth ordering of the detected people. Moreover, using the warped image leads to the removal of the spurious reconstructions. This clearly demonstates the effectiveness of our method.

5. Conclusion

In this work, we aim to improve pose estimation in a multi-person setting. To the best of our knowledge, previous works have not examined the use of optical flow as temporal cues to improve human pose estimation. To this end, we leverage optical flow using RAFT to improve reconstructions at each frame. We propose two schemes for flow fusion on the basis of input type and fusion method. Our experiments demonstrate the effectiveness of using optical flow as a cue for multi-person reconstruction. Our code is available at https://github.com/anirudh-chakravarthy/3D-Project.

6. Future Work

In the future, we aim to explore better methods to leverage temporal information. This could be achieved by using scene flow instead of optical flow or introducing an flow-based supervision in addition to pose-regression losses. While we explore early fusion methods in this work, another future direction could be to explore alternate fusion strategies to better leverage the temporal cues. Finally, due to lack of compute, we were unable to effectively analyze the concatenation strategy. We intend to train these networks from scratch using the SMPL R-CNN training protocol to concretely comment on it's performance.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 4

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 2

[1]The implementation can be found at: https://github.com/ddddwee1/MuPoTS3D-Evaluation

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 2

[4] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2232–2241, 2019. 2

[5] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[6] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[8] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 1, 2, 3, 4, 5, 6

[9] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 1, 2

[10] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 1, 2

[11] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[12] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 1, 2

[13] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 6

[14] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 4

[15] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2

[16] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017. 2

[17] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2

[18] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2329–2336, 2014. 2

[19] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[20] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018. 2

[21] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 2

[22] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6289–6298, 2017. 2

[23] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018. 2

[24] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017. 2

[25] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2821–2840, 2012. 2

[26] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenyà, Carme Torras, and Francesc Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2673–2680. IEEE, 2012. 2

[27] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate

scans from an image in less than a second. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5330–5339, 2019. 2

[28] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018. 1, 2

[29] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3

[30] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3941–3950, 2017. 2

[31] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2500–2509, 2017. 2

[32] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 1, 2

[33] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 2

[34] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017. 2