Multi-person Implicit Reconstruction from a Single Image

Armin Mustafa[1] Akin Caliskan[1] Lourdes Agapito[2] Adrian Hilton[1]
[1]CVSSP, University of Surrey [2]Department of Computer Science, University College London

Abstract

We present a new end-to-end learning framework to obtain detailed and spatially coherent reconstructions of multiple people from a single image. Existing multi-person methods suffer from two main drawbacks: they are often model-based and therefore cannot capture accurate 3D models of people with loose clothing and hair; or they require manual intervention to resolve occlusions or interactions. Our method addresses both limitations by introducing the first end-to-end learning approach to perform model-free implicit reconstruction for realistic 3D capture of multiple clothed people in arbitrary poses (with occlusions) from a single image. Our network simultaneously estimates the 3D geometry of each person and their 6DOF spatial locations, to obtain a coherent multi-human reconstruction. In addition, we introduce a new synthetic dataset that depicts images with a varying number of inter-occluded humans and a variety of clothing and hair styles. We demonstrate robust, high-resolution reconstructions on images of multiple humans with complex occlusions, loose clothing and a large variety of poses and scenes. Our quantitative evaluation on both synthetic and real world datasets demonstrates state-of-the-art performance with significant improvements in the accuracy and completeness of the reconstructions over competing approaches.

1. Introduction

Multi-person human reconstruction from a single image finds application in surveillance; film and entertainment including movie production; generating AR/VR content for complex scenes; and sports broadcast. Reconstruction from a single camera is more practical and lower-cost compared to multi-view, as it does not require a complex setup. Immense progress has been made in estimating 3D human pose and shape from a single image or monocular video in the last five years [3, 46, 35, 11, 13]. Methods can be classified as model-based or model-free. Model-based methods use a parametric human body shape model such as SMPL to reconstruct people from a single image [7, 22, 18] including methods that estimate SMPL with clothing top[25].

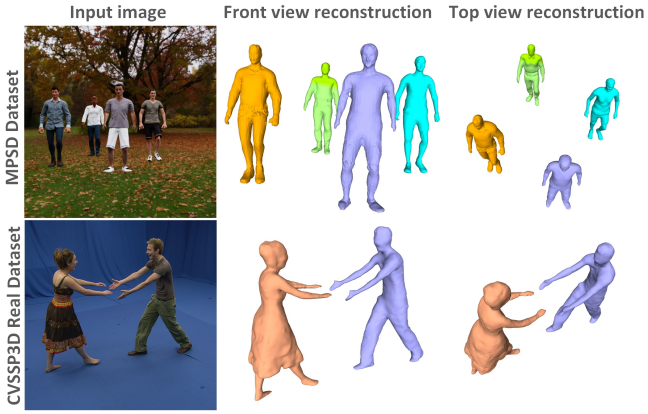

Figure 1. Proposed model-free multi-person spatially coherent implicit reconstruction from a single image with 4 and 2 people on synthetic MPSD and real CVSSP3D dataset.

Model-free methods give a more realistic reconstruction of people with loose clothing and hair details [3, 46, 35, 11]. However all existing methods require an image of a single fully visible person without occlusions to allow reconstruction. Recently, model-based approaches have been introduced that can reconstruct multiple humans in a scene [13, 17, 45] using SMPL, so cannot capture clothing details. In addition, [45] requires manual intervention to mark interaction regions on 3D surfaces to handle inter-person/object occlusions. This paper introduces the first model-free end-to-end approach that reconstructs multiple clothed people from a single image of a crowded scene with inter-person occlusions (see Table 1). Our proposed approach produces a spatially coherent implicit reconstruction of each person together with their 6DOF spatial locations and orientations in the observed scene without any manual intervention, and can handle complex poses, clothing and partial occlusion from a single image, as shown in Fig. 1.

We introduce the first multiple people synthetic dataset and benchmark (MPSD) with realistic image-3D model pairs. MPSD ranges from $2-10$ people per image in a wide variety of clothing, hairstyles and poses with detailed surface geometry and appearance rendered with diverse indoor and outdoor natural backgrounds and realistic scene illumination. This dataset provides the first quantitative benchmark for multi-person single image reconstruction. We pro-

	Model-free	Multi-human	Coherent	Occ.	RGB
[7, 22, 18, 21]	×	×	×	×	✓
[3, 46, 35, 11]	✓	×	×	×	✓
[17, 43, 44]	×	✓	✓	✓	✓
[5, 9]	✓	×	×	✓	×
Holopose [13]	×	✓	×	✓	✓
PHOSA [45]	×	✓	✓	✓	✓
Proposed	✓	✓	✓	✓	✓

Table 1. Comparison of our method with existing 3D shape estimation methods. Occ - Occlusions and Image - Single image

pose an end-to-end method trained on the MPSD dataset that estimates 3D reconstructions of each person and their 6DOF location/orientation by exploiting single image depth and instance segmentation. Example results from the proposed single image multi-human spatially coherent implicit reconstruction are shown in Fig. 1. Our contributions are:

- The first approach for model-free reconstruction of multiple people from a single image with accurate spatial arrangement.
- An end-to-end framework using cascaded multitask networks for simultaneous implicit 3D reconstruction and 6DOF location/orientation estimation exploiting depth and instance segmentation information.
- A multiple person synthetic image/3D dataset of complex multi-person scenes with inter-person occlusions, realistic clothing, hair, poses, scenes and illumination.
- A method that exploits the advantages of volumetric and implicit 3D shape representations for detailed reconstructions of clothed people from a single image.

2. Related Work

This section reviews work in the area of 3D human shape estimation from a single image for a single person and multiple people in the scene, an overview is shown in Table 1.

2.1. Single Person 3D Shape Estimation

Early techniques for single image 3D shape estimation were model-based [12, 7, 18, 21] with most approaches using the SMPL [23]. Methods fit the SMPL model to a single person in an image by using 2D joints [7], silhouettes [22], and 3D joints [18]. 2D Joints and silhouettes were used for 3D shape estimation in [31] with the SMPL-X model [30] and [25, 42, 6] estimate tight fitting clothing on top of the SMPL model. However human shape estimated from these approaches does not represent the actual shape of humans with loose clothing, hair details and large deformations.

Monocular model-free approaches estimate detailed shape of clothed people [38, 46, 3, 11, 35, 10, 9] without a parametric model. [38] uses a voxel representation, [46] uses normals to obtain a discretized volumetric representation, [11] uses front and back depth maps and [3] uses the SMPL model as an initialization to predict model-free 3D human shape. [42] estimates 3D from a RGBD camera and [29] uses multi-view silhouettes to obtain 3D shape from a single image. PiFU regresses an implicit function to determine the occupancy for any given 3D location [35]. This was superseded by PiFUHD for highly detailed 3D reconstruction [36]. PiFU and PiFUHD do not work well for arbitrary poses. This limitation was partially addressed in [15] by using a semantic deformation field. [5, 9] estimates 3D human shape as an implicit function from a sparse point cloud instead of an RGB image, exploiting the SMPL model. However all these methods reconstruct a single person with a limited range of pose and require full visibility without occlusions. [8] exploits multi-views during training to estimate 3D human shape from a single image in a wide variety of poses. This method uses a voxel representation which results in a quantized output mesh with low resolution and less accurate reconstruction.

In this paper, we introduce a method that exploits the advantages of both volumetric voxel and implicit representations by obtaining an initial voxelized mesh of a person which is refined using an implicit function. The proposed method is trained on multiple views [8] to handle partially occluded people in a wide variety of poses.

2.2. Multiple Person 3D Shape Estimation

There are a limited number of methods that predict 3D shape of multiple humans from a single image. [43, 44] were the first methods to perform multi-human spatially coherent reconstruction from a single image exploiting SMPL by using multiple scene constraints to optimize 3D shape [43] and a feed-forward network to estimate pose and shape for multiple people [44]. Jiang et al [17] performed more accurate and robust SMPL-based multi-human reconstruction by directly regressing the SMLP parameters from pixels. Holopose [13] reconstructs multiple people from a single image using Densepose [14], but with a single scale and no spatial layout. Recently [45] introduced an optimization framework for multi-human and object reconstruction using collision and depth information but it required manual intervention to estimate heights of each class and mark the interaction regions on 3D mesh. However all these methods require the SMPL model and suffer from the same limitations as model-based approaches for single humans.

Our paper introduces the first end-to-end model-free method to implicitly reconstruct multiple humans with loose clothing and hair details in a wide variety of poses with inter-person occlusions from a single crowded image, which is a non-trivial task and an unsolved problem in the literature. The proposed end-to-end method has two multitask networks: the first network estimates instance segmentation and depth from the input image; which is exploited in the second network that simultaneously estimates the 3D shape and 6DOF spatial location and orientation of each person to achieve a spatially coherent implicit 3D reconstruction from a single image.

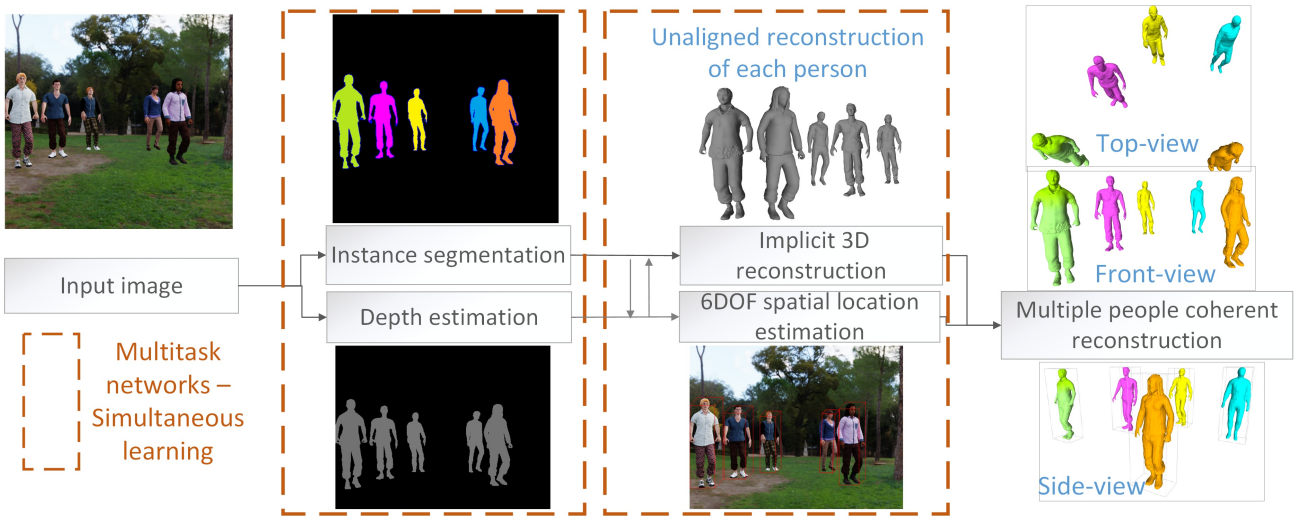

Figure 2. Proposed end-to-end model-free spatially coherent multi-person implicit reconstruction from a single image. The first multitask network estimates segmentation and depth from a single image. The segmentation and depth map is used in the second multitask network to estimate per-instance implicit surface reconstruction of each person together with their 6DOF location and orientation.

3. Methodology

3.1. Overview:

Our proposed learning-based method performs multi-person spatially coherent reconstruction without any manual intervention by first estimating instance segmentation and depth using existing multitask encoder-decoder network [19] from a single image, followed by the second multitask network that estimates implicit 3D reconstruction and 6DOF spatial location of each person simultaneously exploiting the depth and instance segmentation, as shown in Fig. 2. The proposed end-to-end network is trained on a novel *MPSD* dataset introduced in this paper.

Implicit 3D reconstruction, Sec. 3.2, Fig. 3: For each human instance an implicit 3D surface is estimated using an intermediate volumetric voxel-based representation to allow reconstructions from a wide range of poses. The intermediate representation is obtained using multi-view voxel based 3D shape estimation, inspired by [8]. [8] is trained on multiple views for a more accurate reconstruction from a single image. A mutli-view occlusion silhouette loss is added to improve the 3D output [45]. However this gives a quantized voxel output with low resolution. This is addressed in this paper through an additional network that performs implicit function based refinement on the voxel output giving a high-resolution and more complete 3D shape of clothed people exploiting features from voxels, image and depth using hybrid representation learning. The multi-view training in the proposed approach enables us to handle inter-person occlusions and partial visibility in crowded images.

6DOF spatial location/orientation estimation, Sec. 3.4, Fig. 7: The per-person 3D reconstructions from the implicit refinement are unaligned in different coordinate systems leading to spatially incoherent output and incorrect world coordinates. To create a spatially coherent 3D reconstruction from a single image, 6DOF spatial location and orientation is estimated for each person. The instance segmentation and depth is exploited in 6DOF location estimation, inspired by a recent 6DOF pose method [40]. However [40] uses only local depth features, which gives limited performance. We add both local and global depth features along with an ordinal depth loss [17] to the network, to improve the location/orientation estimation in crowded scenes.

The proposed network is trained on a new Multiple People Synthetic Dataset (*MPSD*) with image/3D pairs of varying number of people in the images, 3D ground-truth, instance segmentation and depth maps. This trained network simultaneously predicts 3D shapes and spatial layouts of each person for multi-human 3D reconstruction from a single image with correct relative spatial arrangement.

3.2. Implicit 3D Reconstruction

Existing monocular model-free 3D shape estimation methods either use a volumetric [38] or implicit[35] representation. The volumetric methods work for a wide variety of poses, clothing and hair, however the output 3D is low resolution which lacks surface details. Implicit representation gives a highly detailed surface, but is limited to restricted poses and clothing [15, 8]. In this paper we combine volumetric and implicit representations to benefit from advantages of both, as shown in Fig. 3. In addition our approach handles partial occlusion which none of the previous implicit or volumetric approaches to model-free single human reconstruction allow. The first stage gives an intermediate 3D volumetric voxel representation [8] which handles variation in human poses, clothing and hair and the second stage refines the surface using implicit representation for a

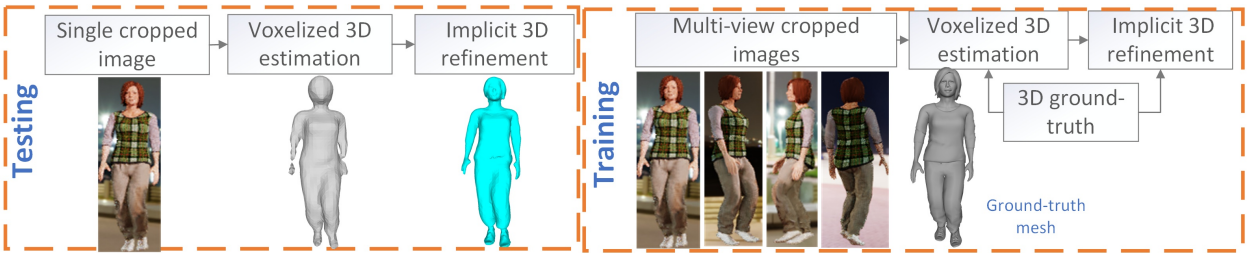

Figure 3. Implicit 3D reconstruction - testing and training framework.

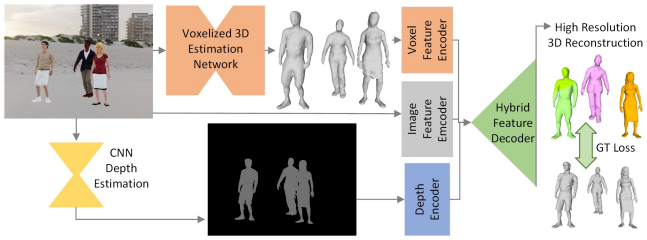

Figure 4. Implicit 3D refinement framework.

more complete high-quality detailed surface reconstruction handling inter-person occlusions. The two stages are:

Voxelized 3D estimation: Intermediate voxelized 3D is obtained using the method [8], which outperforms implicit reconstruction methods [35] in completeness and accuracy of human shape of visible and occluded parts of humans in a wide variety of poses, clothing and hairstyles because of multi-view training. Similarly the proposed method is trained on multiple views for voxelized 3D estimation of each person, as shown in Fig. 3, using the loss function: $L = L_{3D} + \alpha L_M + \beta L_{OS}$. We have added occlusion silhouette loss (L_{OS}) [45] in addition to the 3D ground-truth loss (L_{3D}) and multi-view consistency loss (L_M) [8].

$$L_{3D} = \sum_{p \in V} \sum_{n=1}^{N} \lambda O_p^n \log \widehat{O}_p^n + (1-\lambda)(1-O_p^n)(1-\log \widehat{O}_p^n)$$

$$L_M = \sum_{p \in V} \sum_{\substack{n=l=1 \\ l \neq n}}^{N} \left\| \widehat{O}_p^n - \widehat{O}_p^l \right\|_2 ; L_{OS} = \sum_{n=1}^{N} \sum_{i \in I} S_n^i - m \cdot \hat{S}_n^i$$

where N is the number of views, O_p is the occupancy value of 3D point p in the voxel grid V, $\widehat{O}_p$ is the predicted value, S is the ground-truth silhouette $\hat{S}$ is the rendered silhouette, m is the visibility indicator and α, λ, β are constants chosen experimentally, defined in Sec. 4.1. The multi-view occlusion silhouette loss allows the proposed method to learn 3D shape robust to camera view changes and self-occlusion (Sec. 4.3 for more details). We choose $N = 4$ different views 90° apart for training for a balance between memory use and performance. Multiple people images often suffer from occlusions, multi-view training enables us to address inter-person occlusions by reliably reconstructing the unseen parts of a human using the volumetric representation surface reconstruction. However the surface details are limited due to the relative low resolution of the discrete 3D voxel output. To address this resolution limitation a novel implicit 3D refinement stage is proposed.

Implicit 3D refinement: To obtain high-resolution surface detail from the low resolution 3D voxel reconstruction, an implicit refinement is proposed as shown in Fig. 3. This refinement uses features from the voxelized output, the input image and the predicted depth map, which are fed into the proposed decoder to implicitly compute the occupancy value of an arbitrary 3D point. The proposed decoder takes three inputs: first input is the *voxel feature* extracted from the voxelized 3D of the first stage using the multi-dimensional voxel encoder [9] and the sampled 3D point; second input is the pixel-wise *image feature* extracted from the input RGB image using the hourglass network [35]; and third input is the depth of the corresponding 3D point with respect to the camera view. In the decoding stage, the occupancy value of the 3D point is predicted from the hybrid feature representations. L1 loss is applied between the predicted occupancy ($\widehat{O}$) and the ground-truth occupancy (O) defined as: $L_{GT} = \sum_{p \in \mathscr{M}} \left| O_p - \widehat{O}_p \right|_1$, where p is a 3D point on mesh $\mathscr{M}$. The implicit function for the predicted occupancy is defined as: $\widehat{O} = f(\phi, \varphi, d) \in [0, 1]$, where ϕ are image features, φ are point-wise voxel features and d are depth features. The 3D occupancy values are used to create meshes for each person using Marching cubes [24].The details of the decoder and ablation on the hybrid features are given in Sec. 4. The results of 3D reconstruction before and after *voxel refinement* are illustrated in Fig. 14.

3.3. Multiple People Synthetic Dataset (MPSD)

Existing datasets for training 3D shape estimation methods either have a single person (Surreal[38], THuman [46], 3D-Humans [11], 3DPeople[32], 3DVH[8]) or do not have ground-truth 3D models for multiple people [33, 26]. Training the proposed end-to-end multi-person spatially coherent reconstruction network requires ground-truth 3D human models and their respective 6DOF spatial location/orientation for multiple people. Hence we introduce the first and the largest realistic Multiple People Synthetic Dataset *(MPSD)* generated with synthetic humans in a wide variety of clothing, poses and hair styles in arbitrary positions simulated against different realistic backgrounds. The dataset improves the generalisation of multiple human re-

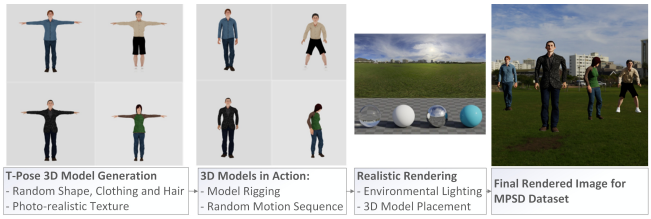

Figure 5. The MPSD dataset generation framework

Figure 6. Examples of multi-person images from the MPSD dataset. The proposed method is able to reconstruct partially occluded people in the scene but not heavily occluded.

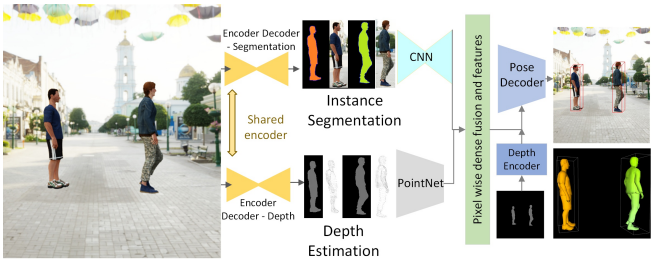

Figure 7. 6DOF spatial location/orientation estimation framework

Figure 8. Comparison of location estimated from the proposed method without the global depth features and ordinal depth loss.

construction in different poses, positions, clothing etc.

The *MPSD* dataset is generated in three steps, as seen in Fig. 5: clothed 3D human model generation; motion sequence application on these models; and multi-view realistic rendering of the models with random placement without intersections. 400 male and female 3D human models with a wide variation in hair, clothing, pose and random positions are generated [1] for various motion sequences [2] for more complete and accurate 3D shape estimation. This dataset consists of varying number of people $\{2, 3, \ldots, 9, 10\}$ at random positions in the scene, as seen in Fig. 6 along with the results from the proposed method. The *MPSD* dataset contains $450k$ *image - 3D models* pairs which are used for single-image multi-human reconstruction. The scene is rendered into 16 camera views with a 512×512 image resolution at each time instant. The *MPSD* dataset provides 3D ground-truth human models, RGB images, depth, instance segmentation and 6DOF spatial locations, which is used to train end-to-and proposed network (Sec. 3.1). The *MPSD* dataset will be released to support research and benchmarking. We will provide RGB images, 6DOF spatial locations, depth maps, instance segmentation and a framework for users to reproduce 3D models, in compliance with the Adobe FUSE licensing terms for release.

3.4. 6DOF Spatial Location/Orientation Estimation

6DOF spatial location/orientations are estimated for each 3D instance, as shown in Fig. 7 to obtain a spatially coherent reconstruction consistent from different views. Previous approaches for multi-human spatially coherent reconstruction either incorporate coherency constraints within the SMPL model estimation [17] or use an optimization framework to estimate the 6DOF pose and scale [45]. In this paper we estimate 6DOF location and orientation for spatially coherent reconstruction as the proposed implicit reconstruction gives shape at the same scale as the input image, removing the requirement of estimating scale.

Out of the many existing 6DOF object pose approaches [40, 34, 41, 37], we choose [40] as our baseline method as it uses segmentation and RGBD information within the network instead of other methods that either use only RGB image and depth separately or use costly post-processing steps, limiting their performances in crowded scenes. We use CNN and PointNet networks from [40] to extract features from segmentation and depth and combine them using a dense fusion network to extract pixel-wise dense feature embedding to estimate pose [40], as shown in Fig. 7. Feature embedding networks and pose decoders are used to estimate the 6DOF location/orientation. In addition to the local depth features [40] we also use global depth features in the pose decoder to improve the location estimation, as seen in Fig. 8. The network is trained using a combination of dense pose loss (L_{DP}) [40] and ordinal depth loss (L_{OD}) [17], $L_{Pose} = L_{DP} + \gamma L_{OD}$ defined as:

$L_{DP} = \frac{1}{U} \sum_{i \in U} \left(\frac{c_i}{Q} \sum_{j \in Q} \left\| (Rx_j + t) + (\hat{R}_i x_j + \hat{t}_i) \right\| \right)$
$- (w \log c_i)$;

$L_{OD} = \sum_{i \in \mathcal{S}} \log \left(1 + \exp(D_{y(i)}(i) + D_{\hat{y}(i)}(i))\right)$

where R, t are ground-truth pose and $\hat{R}, \hat{t}$ are predicted poses, U is randomly sampled dense-pixel features, Q is randomly selected 3D points, c is confidence score for each prediction, w is weight selected empirically and $D()$ is the

Figure 9. Comparison of proposed method against state-of-the-art single image shape estimation methods - GT is ground-truth and DH is DeepHuman. Differences are highlighted in red circle.

depth value. The ordinal depth loss is applied on the entire depth image which is encoded in the form of features and is input to the pose decoder, as seen in the Fig. 7. This allows for a more accurate pose estimation in the case of humans in close proximity, as seen in Fig. 8. The proposed method with global depth encoded features and ordinal depth loss has been evaluated to perform the best.

4. Evaluations and Results

An extensive experimental evaluation is presented on multiple people [45, 20] and single person [35, 36, 8, 46] 3D shape estimation against state-of-the-art methods on the proposed MPSD dataset and publically available single [39, 46] and multi person [28, 27, 4] datasets. We were unable to compare with [17] because of unavailability of code and the datasets they used do not have 3D shape only joints.

4.1. Implementation Details and Architecture

The proposed network is trained on the novel MPSD dataset, during training the two multitask networks are trained separately. The first multitask segmentation and depth network is given segmentation mask and depth map and the second multitask 3D and pose network is given 3D ground-truth and pose for training. For testing only a single image is passed to the network. MPSD is split into training and test sets $(70-30)$ such that for each category with $2, 3...10$ number of people, 30 scenes are used to testing and 70 scenes are used for training. The models in the 30 test scenes are not seen during training for fair evaluation. The constants are: $\alpha = 0.2, \beta = 0.1, \gamma = 0.1, w = 0.001$.

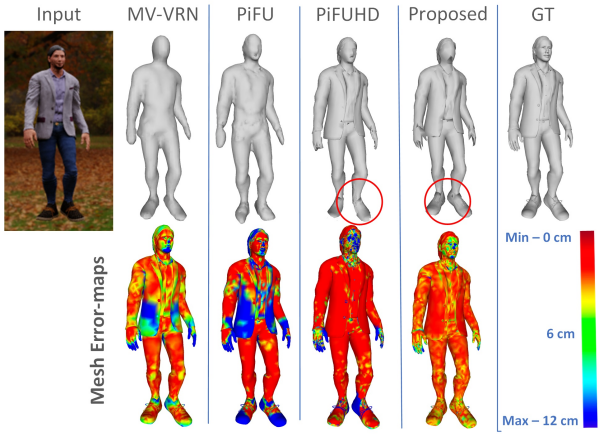

Figure 10. Point to surface (P2S) error maps against ground-truth mesh for proposed method and state-of-the-art methods - GT is ground-truth (error - red - no error and blue - max error).

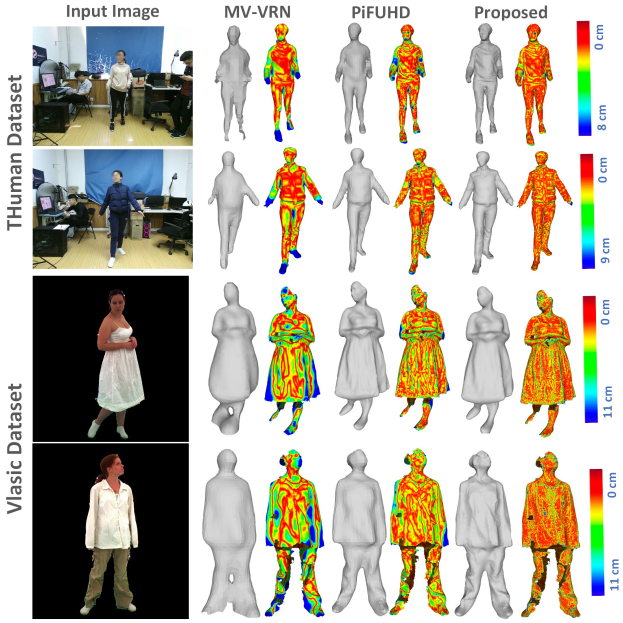

Figure 11. Results and P2S errors against 3D ground-truth for ours, MV-VRN and PiFUHD on real datasets-THuman and Vlasic (error- red - no error and blue - max error)

The implicit refinement decoder network consists of stacked linear 1D convolution layers, which takes concatenated features as input and outputs the 3D occupancy. The voxelized 3D estimation network is similar to MV-VRN[8]. For further network details including the pose decoder network in pose estimation (Sec. 3.4) and more implementation details please refer to the supplementary material.

4.2. Comparative Evaluations

Evaluation on public datasets: The proposed network is trained on MPSD dataset and fine tuned on the training split of the publically available datasets with 3D, given below:
Vlasic [39]: Real single person dataset with 3D
THuman[46]: Real single person dataset with 3D

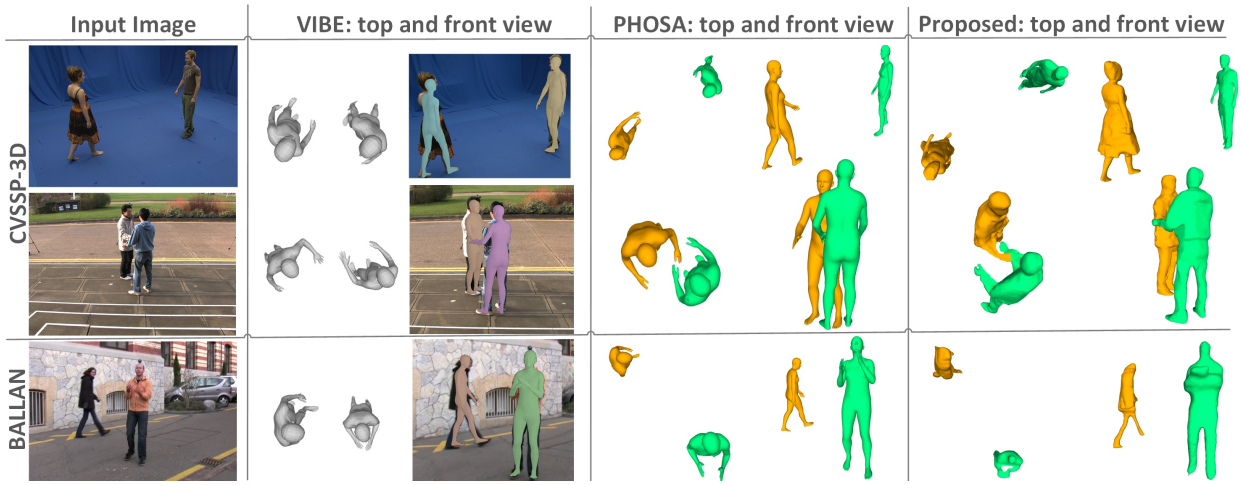

Figure 12. Comparison of proposed multi-human reconstruction against state-of-the-art methods on real datasets - CVSSP3D[28] and BALLAN[4]. PHOSA and proposed method give coherent reconstruction. Vibe does not estimate 3D spatial locations of each person.

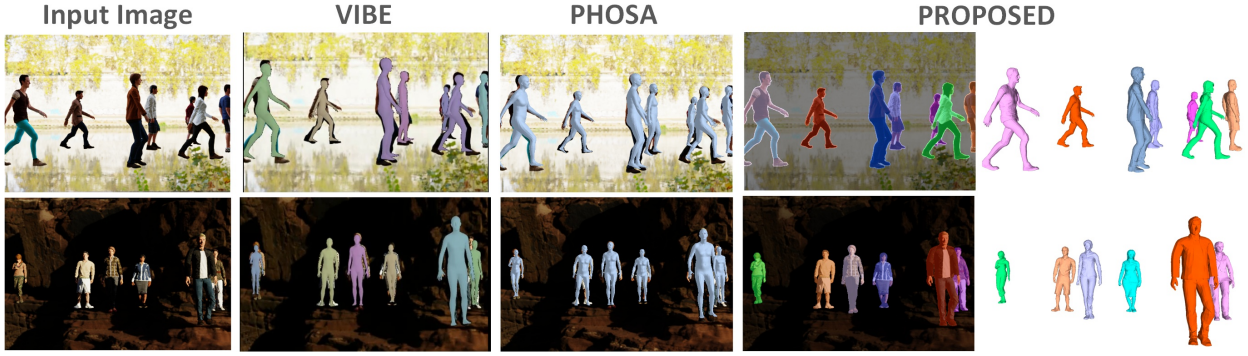

Figure 13. Comparison of mesh projections of single image multi-human reconstruction against the proposed method. Proposed method gives accurate reconstruction with clothing details, unlike PHOSA and Vibe that give SMPL models with no clothing details.

Datasets	MPSD			Vlasic			THuman		
Method	CD	3DIoU	P2S	CD	3DIoU	P2S	CD	3DIoU	P2S
MVVRN	2.11	61%	2.82	2.77	55%	3.39	2.98	53%	3.57
PiFUHD	1.66	68%	1.93	2.14	65%	2.72	2.56	60%	2.97
PiFU	1.98	59%	2.72	2.80	48%	3.22	3.03	45%	3.96
DeepH	3.15	48%	4.02	4.26	43%	5.58	4.67	39%	5.88
Ours	**1.47**	**71%**	**1.88**	**1.93**	**67%**	**2.26**	**2.07**	**65%**	**2.86**

Table 2. Quantitative comparison of our method with state-of-the-art single person methods- CD-Chamfer distance, P2S-Point to surface error, and 3DIoU-3D Intersection of Union. CD and P2S - lower the better, 3DIoU - higher the better. DeepH-DeepHuman

CVSSP3D [28]: Real multiple people dataset with 3D

Ballan [4]: Real multiple people dataset with 3D state-of-the-art single person shape estimation methods.

Single person evaluation on MPSD dataset: We crop the image of one person with no occlusions from proposed the MPSD dataset for fair single person reconstruction evaluation with existing state-of-the-art methods - PiFU [35], PiFUHD [36], MV-VRN [8], and DeepHuman[46]. PiFU and MV-VRN are trained on the MPSD datasets and due to unavailability of code DeepHuman and PiFUHD are only tested on MPSD. Three error metrics are computed using the ground-truth 3D models to measure the quality of shape reconstruction: Chamfer Distance (CD), Point to surface errors (P2S) [36] and 3D Intersection of Union (3D IoU) [16]. Qualitative comparison is shown in Fig. 9 and quantitative evaluation is shown in Fig. 10 and Table 2. The comparison demonstrates that the proposed method gives a high-resolution reconstruction comparable to PiFUHD and significantly better than PiFU, MV-VRN, DeepHuman. The quality and completeness of the reconstruction in the unseen parts of the object are better compared to PiFUHD (side-view in Fig. 9). Results and comparative evaluation on Vlasic and THuman datasets against PiFUHD and MV-VRN in Fig. 11 shows that the proposed method significantly outperforms existing methods on real datasets.

Multi person evaluation on MPSD dataset: The results against multiple people 3D shape estimation methods PHOSA[45] and VIBE[20] are shown in Fig. 12. Quantitative comparison is given in Table 3 against Chamfer distance, P2S and 2D Intersection of Union [28]. Both PHOSA and VIBE estimate the SMPL model of each person in the scene unlike the proposed method which gives model-free realistic reconstruction of people, which align with the object boundaries more closely especially in the case of loose clothing, as seen in Fig. 13. Results on real datasets CVSSP3D and Ballan datasets against PHOSA and

VIBE are illustrated in Fig. 12 and on MPSD dataset is shown in Fig.13, demonstrating that the proposed method outperforms existing methods on both real and synthetic datasets in quality and coherency of the reconstruction. We also compare the 6DOF spatial location and orientation estimated from the proposed method and PHOSA in Table 4 using AUC metric defined in [40] for synthetic and real dataset and the results are comparable.

Datasets	MPSD			CVSSP3D			Ballan		
Method	CD	2DIoU	P2S	CD	2DIoU	P2S	CD	2DIoU	P2S
PHOSA	5.77	73%	6.92	7.73	69%	8.32	9.11	67%	10.04
Vibe	6.12	68%	7.44	8.20	63%	9.14	10.36	64%	11.38
Ours	**1.47**	**84%**	**1.88**	**1.97**	**80%**	**2.31**	**2.27**	**78%**	**3.02**

Table 3. Comparison of the proposed method against multiple people reconstruction methods- CD is Chamfer distance ↓, 2DIoU is 2D Intersection of Union ↑, and P2S is Point to surface error ↓.

	MPSD	CVSSP3D		MPSD	CVSSP3D
PHOSA	75.6	80.4	Proposed	79.8	82.1

Table 4. Comparison of the spatial location and orientation of the proposed method against PHOSA for the AUC metric[40].

	CD ↓	P2S ↓	3DIoU ↑	2DIoU ↑
Proposed	**1.47**	**1.88**	**71%**	**84%**
W/o L_{OS} & implicit	2.11	2.82	61%	72%
W/o Implicit	1.98	2.71	62%	74%
W/o L_{OS}	1.63	2.44	65%	79%
Implicit with 3D & Depth	1.71	2.42	64%	79%
Implicit with 3D & RGB	1.65	2.27	67%	80%
If-net [9]	1.78	2.59	63%	77%

Table 5. Ablation study of the proposed method on MPSD dataset.

4.3. Ablation Study

The abaltion for implicit 3D reconstruction (Sec. 3.2) of the proposed method with and without the implicit refinement is shown in Fig. 14. There is a considerable improvement in the details, completeness and quality of the surface reconstruction. In the implicit refinement encoded features are input from RGB image, depth and 3D voxels. We present ablation without the silhouette loss in the voxelized 3D estimation (L_{OS}), without implicit refinement, with implicit refinement only with 3D and depth features (Implicit-3D & Depth), with implicit refinement with 3D and RGB features (Implicit-3D & RGB) and the proposed method in Table 5 for all error metrics defined before - CD, P2S, 3DIoU and 2DIoU. The proposed methods performs the best in terms of quality and completeness of the 3D shape when all 3 features are encoded in the implicit refinement network (Fig. 4).

We also compare our implicit refinement with If-net [9] in Fig. 15. The input to the implicit refinement network and If-net is the voxelized output from the first stage of the implicit 3D reconstruction (Sec. 3.2). The 3D surface from the proposed method has more details compared to If-net.

Limitations: Like existing methods, the proposed method cannot handle extreme poses (handstand) and heavy occlusions ($> 50\%$). As it's a multi-level approach like existing methods [35], it relies on the success of instance segmentation and depth. Also no temporal information is exploited potentially leading to a temporally incoherent output.

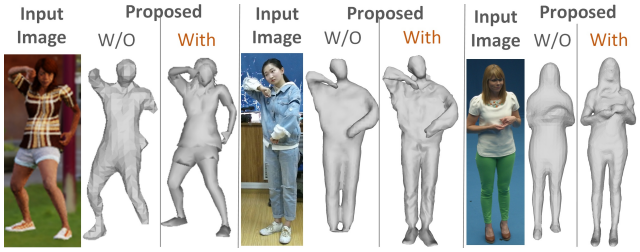

Figure 14. Results of proposed method with and without implicit refinement on synthetic MPSD (left) and real THuman and CVSSP3D (right) datasets.

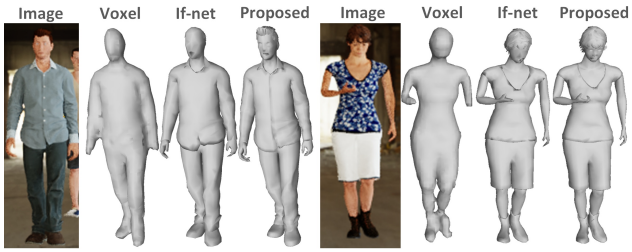

Figure 15. Comparison of proposed approach with If-net [9]

5. Conclusion

This paper presents an end-to-end learning framework for spatially coherent model-free implicit reconstruction of a multi-person single image. The method gives realistic implicit reconstructions of people with loose clothing and hair without any manual intervention compared to previous SMPL based approaches and works with partially occluded people. The proposed methodology combines the advantages of explicit volumetric representations to reconstruct a wide range of poses and implicit function representations for reconstruction of shape detail. A multitask network is used to simultaneously estimate 3D shape and 6DOF spatial location and orientation of each person in the image for a spatially coherent multi-human reconstruction. The proposed network is trained on our novel *MPSD* dataset with image-3D ground-truth pairs. High-resolution, robust and spatially coherent reconstructions are demonstrated on synthetic and real datasets on complex multi-person single images with partial occlusions, in a variety of clothing, poses, and scenes. Extensive comparative evaluation with single person 3D shape estimation methods demonstrates improvement in the accuracy, completeness and detail of the reconstruction. Comparative evaluation with state-of-the-art multi-human reconstruction methods shows that the proposed approach achieves a significantly better reconstruction of clothed humans as the existing methods only give SMPL model reconstruction of each person in the scene.

Acknowledgement: Supported by Dr Mustafa's RAEng Fellowship and EPSRC Platform Grant EP/P022529/1.

References

[1] Adobe. Fuse, https://www.adobe.com/products/fuse.html, 2020. 5

[2] Adobe. Mixamo, https://www.mixamo.com/, 2020. 5

[3] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[4] Luca Ballan, Gabriel J. Brostow, Jens Puwein, and Marc Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2010)*, pages 1–11, July 2010. 6, 7

[5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*, August 2020. 2

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2

[7] 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 (ECCV)*, 2016. 1, 2

[8] Akin Caliskan, Armin Mustafa, Evren Imre, and Adrian Hilton. Multi-view consistency loss for improved single-image 3d reconstruction of clothed people. In *Asian Conference on Computer Vision (ACCV)*, 2020. 2, 3, 4, 6, 7

[9] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 4, 8

[10] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[11] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4

[12] Peng Guan, Alexander Weiss, Alexandru O. Bălan, and Michael J. Black. Estimating human shape and pose from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1381–1388, 2009. 2

[13] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 2

[15] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3093–3102, 2020. 2, 3

[16] Krishna Murthy Jatavallabhula, Edward Smith, Jean-Francois Lafleche, Clement Fuji Tsang, Artem Rozantsev, Wenzheng Chen, Tommy Xiang, Rev Lebaredian, and Sanja Fidler. Kaolin: A pytorch library for accelerating 3d deep learning research, 2019. 7

[17] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 5, 6

[18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[19] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. 6, 7

[21] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[22] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions in Graphics*, 34(6), Oct. 2015. 2

[24] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *COMPUTER GRAPHICS*, 21(4):163–169, 1987. 4

[25] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[26] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *ACM Transactions on Graphics*, 39(4), July 2020. 4

[27] A. Mustafa, H. Kim, and A. Hilton. Semantically coherent 4d scene flow of dynamic scenes. *International Journal of Computer Vision*, 2019. 6

[28] Armin Mustafa, Chris Russell, and Adrian Hilton. U4d: Unsupervised 4d dynamic scene understanding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6, 7

[29] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. SiCloPe: Silhouette-Based Clothed People. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11, June 2019. 2

[30] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, June 2019. 2

[31] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 06 2018. 2

[32] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4

[33] Anurag Ranjan, David T. Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J. Black. Learning multi-human optical flow. *International Journal of Computer Vision (IJCV)*, Jan. 2020. 4

[34] Martin Runz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, and Richard Newcombe. Frodo: From detections to 3d objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[35] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 3, 4, 6, 7, 8

[36] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6, 7

[37] Meng Tian, Liang Pan, Marcelo H Ang Jr, and Gim Hee Lee. Robust 6d object pose estimation by learning rgb-d features. In *International Conference on Robotics and Automation (ICRA)*, 2020. 5

[38] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 4

[39] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM Transactions on Graphics*, 2009. 6

[40] C. Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, L. Fei-Fei, and S. Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3338–3347, 2019. 3, 5, 8

[41] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems (RSS)*, 2018. 5

[42] Tao Yu, Zerong Zheng, Y. Zhong, Jianhui Zhao, Q. Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap : Single-view human performance capture with cloth simulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5499–5509, 2019. 2

[43] A. Zanfir, Elisabeta Marinoiu, and C. Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 2

[44] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*, volume 31, pages 8410–8419, 2018. 2

[45] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 4, 5, 6, 7

[46] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4, 6, 7