Learning to Regress Bodies from Images using Differentiable Semantic Rendering

Sai Kumar Dwivedi[1] Nikos Athanasiou[1] Muhammed Kocabas[1,2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zurich
{sdwivedi, nathanasiou, mkocabas, black}@tue.mpg.de

Figure 1: **Differentiable Semantic Rendering (DSR)**. A state-of-the-art approach [14] (purple) fails to estimate accurate 3D pose and shape for in-the-wild scenarios. We address this by exploiting the clothing semantics of the human body. Our approach, DSR, (blue) captures more accurate 3D pose and shape compared to previous work.

Abstract

*Learning to regress 3D human body shape and pose (e.g. SMPL parameters) from monocular images typically exploits losses on 2D keypoints, silhouettes, and/or part-segmentation when 3D training data is not available. Such losses, however, are limited because 2D keypoints do not supervise body shape and segmentations of people in clothing do not match projected minimally-clothed SMPL shapes. To exploit richer image information about clothed people, we introduce higher-level semantic information about clothing to penalize clothed and non-clothed regions of the human body differently. To do so, we train a body regressor using a novel "**D**ifferentiable **S**emantic **R**endering **(DSR)**" loss. For Minimally-Clothed (MC) regions, we define the DSR-MC loss, which encourages a tight match between a rendered SMPL body and the minimally-clothed regions of the image. For clothed regions, we define the DSR-C loss to encourage the rendered SMPL body to be inside the clothing mask. To ensure end-to-end differentiable training, we learn a semantic clothing prior for SMPL vertices from thousands of clothed human scans. We perform extensive qualitative and quantitative experiments to evaluate the role of clothing semantics on the accuracy of 3D human pose and shape estimation. We outperform all previous state-of-the-art methods on 3DPW and Human3.6M and obtain on par results on MPI-INF-3DHP. Code and trained models are available for research at https://dsr.is.tue.mpg.de/.*

1. Introduction

Estimating 3D human pose and shape from in-the-wild images has received great research interest [5, 14, 15, 18, 20, 30, 34, 54] because of its varied applications in animation, games, and the fashion industry. One aspect that makes this problem challenging is the difficulty of obtaining accurate 3D ground-truth annotations, as they require either specialized –mostly indoors– MoCap systems or careful calibration and setup of IMU sensors [46]. Such data would facilitate training robust regressors paving the way for estimating human-scene interaction with greater granularity.

Given the lack of in-the-wild 3D ground-truth, the vast majority of previous methods focus on 2D keypoints [5, 14] with some learned 3D priors. Even though sparse 2D keypoints give useful constrained, relying only on these leads to unrealistic poses because of depth ambiguities and occlu-

sion. They also do not provide reliable information about body shape. On the other hand, relying too strongly on 3D priors introduces bias. To circumvent this problem, recent approaches [30, 34, 50] propose to use part-segmentations or silhouettes. However, there is a mismatch between part-segmentations/silhouettes and projected SMPL bodies since segmentation covers clothed bodies while the common 3D body models are minimally clothed. We propose an alternative approach to compensate for limited 3D supervision that leverages high-level 2D image cues.

Specifically, we propose more detailed clothing segmentation labels to supervise a neural network. Traditional multi-class clothing segmentation approaches cannot be directly applied as the segmentation loss tries to exactly match the rendered SMPL body. Hence, to make use of such labels, we need to reason about which parts of the SMPL body model correspond to which clothing label. This is non-trivial to obtain because a body part can be covered by many clothing types. Therefore, we learn a semantic clothing prior from a large-scale clothed human scan dataset, which has varied subjects, poses and camera views to which the SMPL body is fitted [31]. This prior encodes the likelihood of clothing types given a vertex on the SMPL body model, which gives the correspondence between segmentation labels and the SMPL body surface. Then, we use this prior to calculate a loss between the SMPL body and observed clothing labels in images. To achieve this we introduce ***D***ifferentiable ***S***emantic ***R***endering (***DSR***), a novel loss that supervises the training of 3D body regression with clothing semantics using weak supervision [8].

Our novel loss has two components: *DSR-C* for supervising the clothing region and *DSR-MC* for the minimally-clothed region. A high-level illustration of our idea is shown in Fig. 2. While the former ensures that the rendered SMPL mesh stays inside the observed clothing label, the latter tries to tightly match the rendered SMPL mesh to the 2D minimal-clothing mask. The loss between the rendered output and the target mask is back-propagated using a differentiable renderer. Specifically, for the *DSR-MC* term, we apply pixel-level supervision for tight-fitting with the minimal-clothing regions, while for *DSR-C*, we minimize the negative log probability of a SMPL semantic part label being inside the respective segmentation mask. For example, there will be a high penalty if the rendered vertices with a high probability of being "shirt" fall in the "pants" segmentation pixels. To ensure that our method is fast and differentiable, we render the semantic class probabilities computed from 3D scans as textures of the SMPL mesh.

While training, DSR can be used as an additional loss in any neural network-based human body estimator that predicts SMPL parameters. First, we examine the effect of our approach over a baseline full-body mask supervision and 3D joint only supervision which verifies our hypothesis about the value of clothing semantics. Then, we perform extensive comparisons and show that DSR outperforms previous state-of-the-art methods as shown in Fig. 1. In summary:

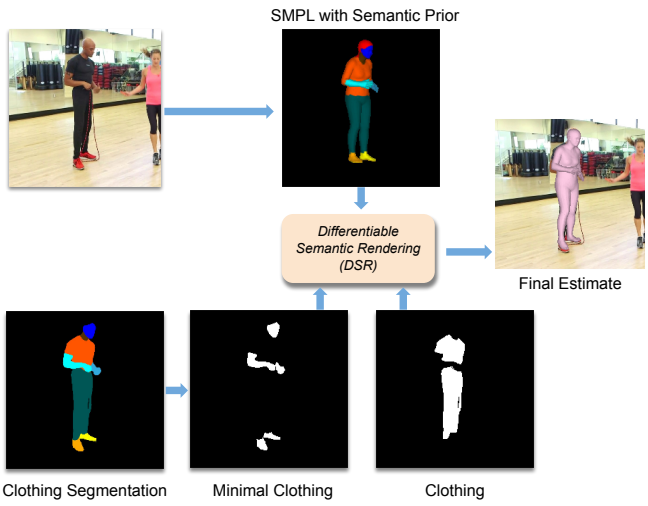

Figure 2: **DSR Idea** - For more accurate human body estimation, we supervise 3D body regression training with clothed and minimal-clothed regions differently using our novel *DSR* loss and our learned semantic prior. The semantic prior represents a distribution over possible clothing labels for each vertex. For easier illustration, we depict the most likely labels per-vertex here.

1. We explore the importance of clothing semantics for 3D human body estimation by introducing a novel differentiable semantic rendering loss that distinguishes between clothed and minimally-clothed regions.
2. We estimate a semantic clothing prior for SMPL from 3D scans of clothed people for our method which can be used also for other cases when a vertex clothing probability for a 3D SMPL body is required.
3. We outperform all state-of-the-art methods on 3DPW and Human3.6M and obtain on par results on MPI-INF-3DHP, suggesting the value of using human parsing and semantics for more accurate human body estimation.

2. Related Work

Estimating human pose and shape is a vastly growing field using different sources of supervision and input (image, video, keypoints, etc.). Here, we focus on different works that estimate 3D human pose and shape from an RGB image. We refer to recent surveys [7, 38] for more details.

2.1. Image cues and 2D/3D joints

Towards estimating 3D human pose and shape, initial attempts focus on estimating the coordinates of 3D joints or heatmaps [12, 21, 33, 39–41, 43] from images using geometric assumptions for the human body and 3D training data. However, those approaches require 3D ground-truth data, which are limited in terms of pose variation, quantity and background, and lack generalization to in-the-wild images. The vast progress in 2D pose detection [6, 29, 42, 49], along with the introduction of parametric body models of pose and shape [3, 24] lead to significant progress and high-quality in-the-wild 3D humans. In [5] the authors use 2D keypoints to obtain SMPL parameters with an optimization-based approach, while this process improves via human annotations on predicted fits [20]. Martinez et al. [27] show that lifting the predictions of a 2D keypoint detector provides a reasonable baseline for the 3D pose. Pavlakos et al. [32] use additional ordinal depth annotations for weak 3D supervision. Kolotouros et al. [19] regress vertex locations using a sub-sampled SMPL mesh and Graph-CNNs. Xiang et al. [47] extract joint confidence maps and 3D orientation information via CNNs and pair them with a deformable body model. Furthermore, in HMR [15], a regressor from 2D joints to SMPL parameters is trained, using a discriminator with unpaired 3D data [26] to encourage plausible poses. Along these lines, some recent approaches using video as input, have applied similar methods to predict temporal kinematics of 3D bodies [16] and estimate the body using temporal features and a motion discriminator [17]. Another approach [51], uses a disentanglement of the skeleton from the 3D human mesh paired with a self-attention network to ensure temporal coherence. SPIN [18], revisits optimizations methods in collaboration with neural networks as it uses a network [15] that provides an initial estimate to the optimization process (SMPLify). Moreover, a regressor-based alternative suggests the use of the 3D neural regressor as a pose prior [14]. Although such methods produce promising results, they typically estimate average body shapes, are not robust to occlusion, and produce poses that are only approximate. Without 3D training data the problem is hard.

2.2. Image alignment and pixel-level supervision

Concurrently, there is a line of research that uses additional constraints, in addition to image features and 2D/3D joints, to better align the body with the image such as dense body landmarks, silhouettes, body part segmentation or pixel aligned implicit functions. Initial seminal lines of work, use a few keypoints along with the SCAPE body model and optimize 3D body shape with silhouettes and smooth shading [9]. Along these lines, Balan et al. [4] propose a distance function for the connected silhouette to ensure the rendered 3D model falls inside the mask. Later work uses 2D keypoints, background segmentation, and SMPL to extract 3D bodies from images [44], similar to [34] who use silhouettes for supervision. Even silhouettes, although they provide supervision when keypoints fail, are often ambiguous in the case of self-occlusion. Towards a detailed alignment of the 3D human body surface and pixels the authors in [10] introduce a dataset with image-to-surface correspondence from MS-COCO [22] and a variant of Mask-RCNN that regresses UV coordinates from images. Part-segmentation masks and IUV are also used in [30] and [54], respectively, as dense supervision. A continuous UV map of SMPL for direct pixel correspondence of the image and the 3D mesh is introduced in [54]. In a similar approach [50], exploits IUV maps as a proxy representation. It estimates SMPL parameters by minimising dense body landmarks and human part masks and also by using motion discriminator. While a large majority of the aforementioned work leverages a parametric 3D body model, there is some recent work that uses voxel representations along with 2D pose and part segmentation supervision [45] or employs implicit functions with surface reconstruction techniques to reconstruct clothed humans. Although these approaches output fine-grain details, they are unable to capture the shape under clothing and are prone to occlusion [11, 36, 37]. An interesting approach is proposed in [55] where a partial UV map of the person is used and the human pose estimation is formulated as an image inpainting problem. Another work that explores scene semantics [35], predicts the label of an occluding object and employs this information to detect invisible joints. Finally, Zanfir et al. [52] represent the body with a normalizing flows-based latent space and use body part segmentation supervision to estimate 3D human body pose from videos and images, unifying different previous approaches. Clothing segmentation is used in [48] for clothing deformation to penalize the vertex offset of the clothed body if the rendered vertex falls outside the clothing boundary.

Most of these approaches are based on joints, silhouettes and part-segmentation masks using approximate supervision for the pose of a person in clothing. We claim that there is more that an image can tell us about human pose. Our key insight is that clothing for different parts of the body conveys important information for detailed fitting. We employ an off-the-shelf 2D semantic segmentation method [8] and a semantic clothing prior to apply these labels in 3D. Given those, we supervise clothed and minimal-clothed regions separately, yielding more aligned fits of 3D humans.

3. Method

DSR uses high-level semantic information for more accurate pose and shape estimation using two additional loss terms DSR-MC and DSR-C, as shown in Figure 3. DSR takes an image I as input which passes through a CNN.

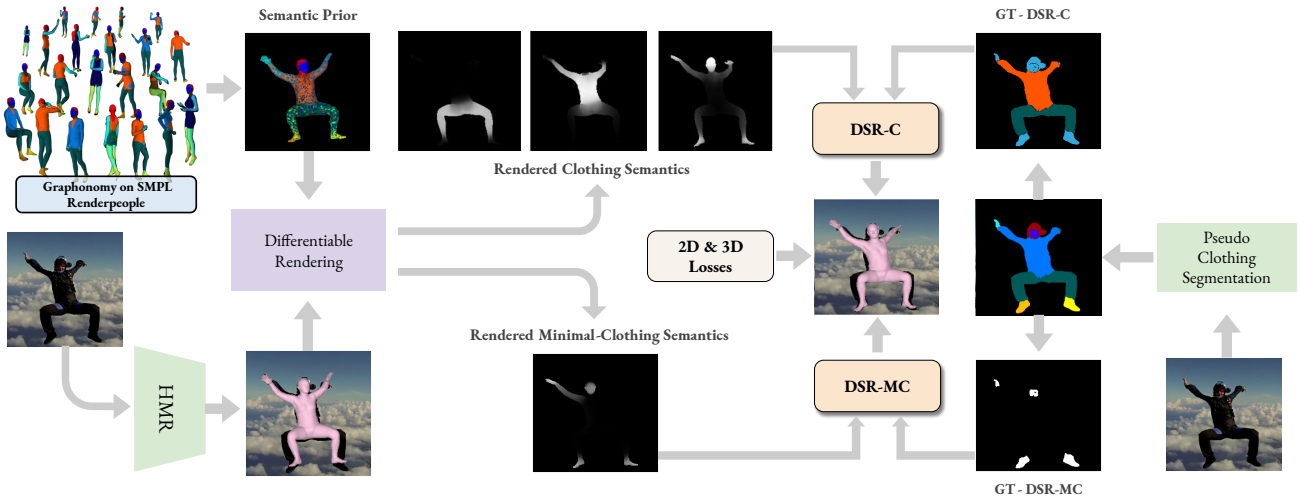

Figure 3: **Illustration of DSR** - SMPL is rendered with the *semantic prior* learned from RenderPeople scans. The two novel loss terms are calculated based on different semantic regions of the clothed person. DSR-MC tightly fits the minimal-clothed region, while DSR-C ensures that the rendered body lies within the clothing boundaries.

Then, the image features $\Phi(I)$ are fed to an iterative regressor, similar to HMR [15], to estimate the parameters of SMPL body model, $\hat{\Theta}$. Given the rendered SMPL mesh, we apply our novel DSR-MC and DSR-C losses in addition to standard loss terms used in EFT [14]. SMPL is a parametric body model that represents body pose and shape by $\Theta = [\theta \in \mathbb{R}^{72}, \beta \in \mathbb{R}^{10}]$. The pose parameters θ include the global rotation and rotations of 23 body joints in axis-angle format and the shape parameters β consist of the first 10 coefficients of a PCA shape space. SMPL model is a differentiable function $\mathcal{M}(\theta, \beta) \in \mathbb{R}^{6890\times 3}$ that outputs a 3D mesh according to the pose and shape parameters.

Clothing Semantic Information. Ground-truth clothing segmentations are expensive to obtain for in-the-wild datasets, which limits the scalability of such an approach. Hence, to analyze the importance of clothing semantics for human pose and shape estimation, we employ an off-the-shelf segmentation model to generate pseudo ground-truth clothing semantics. Graphonomy [8] is a state-of-the-art clothing segmentation model that uses inter and intra graph transfer learning for unifying different clothing datasets and produces 20 clothing labels and body part segmentations. As DSR-MC reasons about the minimal-clothing region, we use a binary mask comprised of 5 labels - *LeftArm, RightArm, LeftShoe, RightShoe and Face* from Graphonomy as ground-truth (whenever available). For DSR-C, we use 4 labels - *UpperClothes, LowerClothes, Minimal-Clothing and Background*. We run the *Universal Model* of Graphonomy on all the datasets to generate pseudo-truth clothing segmentations. For more details on the generation of pseudo-ground truth, cleaning of obtained masks and mapping of graphonomy labels for DSR-C and DSR-MC, please refer to the Sup. Mat.

Semantic Prior for SMPL. To use the semantic information obtained from Graphonomy as pseudo ground-truth training labels, we need a semantic prior of clothing for SMPL 3D bodies. To achieve this, we use thousands of scans from Renderpeople [1] with varied clothing, subject, pose and 10 camera views for which we have ground-truth SMPL fits from AGORA [31]. We run the universal model of Graphonomy on the rendered images of the scan with 20 clothing and body part segmentation labels. Next, we use the ground-truth SMPL mesh to compute the visible face triangles given the mesh and camera parameters. Then, each visible triangular face is assigned the corresponding segmentation label. We repeat this process for all the available scans. We compute the probability of each vertex being a particular label out of the 20 labels from Graphonomy. This probabilistic label for each vertex is referred to as *semantic prior*. For more details refer to Sup. Mat.

Differentiable Semantic Rendering. We use SoftRas [23] as the differentiable renderer to supervise the estimation of the 3D parametric model using semantic information. It uses a differentiable aggregation process for rendering, which fuses the probabilistic contributions of all mesh triangles with respect to rendered pixels. The semantic prior obtained from AGORA [31] is used as a texture. Specifically, for each semantic label, we render the probability of that label for each visible vertex. Once the semantic probability is rendered as images by SoftRas, the loss is computed on the 2D image output by comparing with the semantic image segmentation and this is backpropagated to change the vertices, in turn, changing the network to give more accurate SMPL parameters.

Standard Losses. As we use the EFT [14] data for training, we use the standard supervision loss $\mathcal{L}_{SD}$ similar to EFT which is defined as:

$$\mathcal{L}_{2D}(\Pi(\mathcal{M}(\hat{\Theta})), j) + \mathcal{L}_{3D}(\mathcal{M}(\hat{\Theta}), J) + \mathcal{L}_{\Theta}(\Theta, \hat{\Theta}) \quad (1)$$

where, $\hat{\Theta}$ are the estimated SMPL parameters, $\mathcal{L}_{2D}$ is the joint reprojection loss, $\mathcal{L}_{3D}$ and $\mathcal{L}_{\Theta}$ are losses on 3D joints and SMPL parameters, respectively. Ground truth 2D joints are represented by j, 3D joints by J, SMPL parameters by Θ and the camera projection function by Π.

DSR - Minimal-Clothing. For minimal-clothing, we choose five labels from Graphonomy namely, *LeftArm, RightArm, LeftShoe, RightShoe and Face*, which often appear similar in shape to the rendered SMPL body; i.e. look roughly "naked." For a particular image, we take the clothing segmentation mask given by Graphonomy and create a binary mask G comprising of the valid labels for that image from the available five labels. This forms the ground-truth for DSR-MC denoted by *GT - DSR-MC* in Fig. 3. We render the probability distribution of vertex labels for SMPL precomputed from RenderPeople as textures; these are shown as *Rendered Semantics* in Fig. 3. We only take the probability distribution of vertices that are visible and set the others as zero. Thus, we define the DSR-MC loss to tightly match the corresponding rendered minimal-clothing region of SMPL to the available semantic binary mask as shown in Fig. 3 (bottom).

We study two variants of the loss for DSR-MC: soft-DistM and soft-IOU. Soft-DistM is inspired by the DistM loss of Naked Truth [4] which was originally proposed for estimating body shape under-clothing. Since we render the semantic probability instead of silhouettes, we call it soft-DistM. It is a distance measure function that takes the rendered image R and target binary Graphonomy mask G and is defined as:

$$\mathcal{L}_{MC-sDistM} = \sum_{i,j}(R_{i,j} \cdot d_{i,j}(G))/(\sum_{i,j} R_{i,j})^{3/2} \quad (2)$$

where $R_{i,j}$ are the pixels inside rendered human body and $d_{i,j}$ is a distance function which is zero if pixel (i, j) is inside G. For points outside, it is defined as the Euclidean distance to the closest point on the boundary of G. Soft-DistM can pull the output inside the target because of the sharp difference in penalization between pixels inside the mask and pixels outside. Given a good initial estimate, the Soft-DistM loss ignores spurious and scattered labels outside the region of interest because the loss is high for pixels far away. This is particularly helpful, when using an off-the-shelf segmentation model without instance segmentation, which can give the wrong output for hard examples.

However, soft-DistM cannot fully ensure that the rendered output exactly matches the target as it gives the same penalty for outputs with different percentages of overlap when it is inside the boundary. Hence, we studied soft-IoU, which ensures tight fitting and is calculated as:

$$\mathcal{L}_{MC-sIOU} = \frac{1}{N}\frac{\sum_{(i,j)} P_{i,j} \cdot G_{i,j}}{\sum_{(i,j)} P_{i,j} + G_{i,j} - P_{i,j} \cdot G_{i,j}} \quad (3)$$

where $P_{i,j}$ is the rendered vertex probability at pixel (i, j), $G_{i,j}$ is the graphonomy label for that pixel. Soft-IoU suffers from spurious and scattered labels outside the region of interest and also suffers from the lack of instance segmentation in the off-the-shelf model. However, we choose soft-IoU for the metric for DSR-MC due to better quantitative results in the baseline experiments in Table 1.

DSR - Clothing. The rendered SMPL body mesh cannot exactly match all the target pixels for the clothing region. Hence, for a more accurate estimate of the 3D body model, we want to encourage the rendered SMPL mesh to stay inside the clothing mask. Previous methods [4] define a distance function to deal with such scenarios. However, we have higher-level semantic information than a silhouette to better address this. We have additional boundaries other than the body outline to enforce that a particular semantic part of the SMPL mesh should fall inside the corresponding semantic part of the segmentation mask. Clothing segmentation provides additional boundaries, such as between the upper and lower body or between clothing and skin.

Specifically, we define four labels, *UpperClothes, LowerClothes, MinimalClothing and Background*, shown as four color masks in Fig. 3 (top). We introduce a *MinimalClothing* label for DSR-C to avoid confusion between the background and minimal-clothing region. Without it, the DSR-C loss would give the same penalty when the minimal-clothing region falls on the corresponding target region or the background. As the semantic prior learned from RenderPeople has 20 probability labels per vertex, we add all the probabilities of upper body clothing labels for *Upperclothes*, lower body clothing labels for *LowerClothes* and body part segmentation labels for *MinimalClothing*. We define DSR-C loss as the negative log-likelihood (NLL) of the rendered probability distribution of each vertex belonging to one of the four labels. The rendered probability distribution is first sent through log softmax before applying NLL loss for numerical stability. So, $\mathcal{L}_{DSR-C}$ is defined as

$$\mathcal{L}_{DSR-C} = \sum_{i=1}^{W}\sum_{j=1}^{H} -log(y_{i,j}) \quad (4)$$

where $y_{i,j}$ is the probability output for the vertex at pixel (i, j), H is the height and W is the width of the image. Hence, the total loss $\mathcal{L}_{total} = \mathcal{L}_{SD} + \mathcal{L}_{MC-sIOU} + \mathcal{L}_{DSR-C}$.

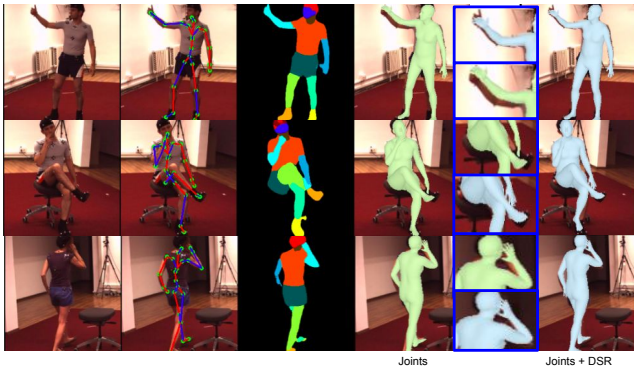

Figure 4: **Are 3D joints enough?** We over-fit a batch of H36M samples on ground-truth (GT) joints (green) and joints with DSR (blue). The weak supervision with semantic information improves accuracy.

4. Experimental Setup

Training Procedure. Following EFT [14], we train a regressor similar to HMR [15] with mixed 3D and 2D datasets. We use the pseudo-ground 3D annotations for 2D datasets from EFT. For 2D data, we only use COCO [22] as including other in-the-wild datasets did not give a performance gain and for 3D datasets, we use Human3.6M [13] and MPI-INF-3DHP [28]. We also use the 3DPW [46] training set for fair comparisons and the same data ratio for mixed 2D and 3D datasets as EFT. For baseline and ablation experiments, we train only on COCO-EFT [14]. For faster training, we initialize the network with SPIN pre-trained weights and use the same hyper-parameters as SPIN [18] and train the model for 100K iterations.

Evaluation Procedure. For state-of-the-art comparisons, we use 3DPW [21], Human3.6M [13] and MPI-INF-3DHP [28]. As in prior work [14, 18], we use the gender information for ground truth meshes on 3DPW. We report results with and without 3DPW training on Procrustes-aligned mean per joint position error (PA-MPJPE), mean per joint position error (MPJPE) and Per Vertex Error (PVE).

Differentiable Semantic Rendering. We use SoftRas [23] to render the probability distribution for DSR-C and DSR-MC. For SoftRas, we use a higher gamma value of 1.0×10^{-1} to ensure the loss affects the occluded part of the body and a lower sigma value of 1.0×10^{-5} to ensure the error does not significantly affect the spatial region. For more details, refer to SoftRas [23]. We render the probability distribution of each triangle face as textures and compute the loss on the RGB channel of the rendered output. We render 5 images for each sample in a batched manner: 1 for DSR-MC and 4 for DSR-C. However, the loss is calculated per individual sample to avoid calculating for samples that do not have a valid segmentation mask. In such cases, the loss is set to zero. After using the heuristics to clean the mask, a valid label set is created for DSR-C and DSR-MC. The weighting parameters for both the components are set to 0.01. As DSR depends on weak supervision from off the shelf clothing segmentation model and hence not robust for hard examples, we enable the loss after 10K iterations into our training.

Method	PAMPJPE ↓	MPJPE ↓	PVE ↓
C-EFT	58.5	101.0	119.3
+ DSR-FB	59.8	102.1	120.3
+ DSR-FB (s-DistM)	58.0	100.2	117.8
+ DSR-MC (s-DistM)	58.2	100.6	118.5
+ DSR-MC (s-IoU)	58.0	100.3	118.1
+ DSR-C	57.6	99.8	117.6
+ DSR-MVP	58.1	100.3	117.8
+ DSR-C + DSR-MC (Ours)	**57.2**	**99.2**	**116.3**

Table 1: **Baseline Comparisons for DSR on 3DPW.** C-EFT is the regressor trained with COCO-EFT and standard losses. DSR-FB is supervised with a full-body silhouette. DSR-MC is minimal-clothing, DSR-C is clothing and DSR-MVP is manual labelling of clothing and minimal-clothing.

5. Results

5.1. Baseline Comparison and Ablation Studies

We perform baseline experiments to (1) motivate the use of semantic rendering and (2) study how the different terms and design choices contribute to the final result as shown in Table 1. As a baseline, we use an HMR [15] based regressor trained on EFT-COCO [14] data and report results on 3DPW (C-EFT). Then, we supervise the baseline with an additional full-body silhouette (DSR-FB) which is a per pixel binary classification loss guided by differentiable rendering. The results deteriorate as the rendered SMPL body does not match the full body. We further train DSR-FB with the Dist-M loss in contrast to per-pixel classification to ensure all body parts (irrespective of clothing) stay inside the silhouette. The result in Table 1 shows that explicit supervision with clothing semantics (Ours) outperforms the naive cloth-agnostic approach. We study the importance of estimating clothing semantics from scans in contrast to manual vertex painting (MVP) of semantic labels as the former gives a distribution over possible clothing labels (20) for each vertex whereas the latter would give only 1. To quantitatively verify the benefit of the probabilistic clothing semantic prior, we take the most likely label per vertex (Fig. 2) as a proxy for MVP. Since we have 1 label per vertex, we use IoU instead of s-IoU. Table 1 shows low performance of a fixed semantic prior (MVP) compared to a probabilistic one (Ours). We also study the individual contribution of DSR-C and DSR-MC on the overall performance and find that the clothing term helps more than the minimal-clothing

Models	3DPW			Human3.6M		MPI-INF-3DHP	
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓
HMR [15]	76.7	130.0	-	56.8	88	89.8	124.2
NBF [30]	-	-	-	59.9	-	-	-
Pavlakos *et al*. [34]	-	-	-	75.9	-	-	-
CMR [19]	70.2	-	-	50.1	-	-	-
SPIN [18]	59.2	96.9	116.4	41.1	62.5	67.5	**105.2**
EFT [14]	54.2	-	-	43.7	-	68.0	-
Zanfir et. al [52] (w/ 3DPW train)	57.1	90.0	-	-	-	-	-
EFT [14] (w/ 3DPW train)	52.2	-	-	43.8	-	67.0	-
DSR	54.1	91.7	105.8	**40.3**	**60.9**	**66.7**	105.3
DSR (w/ 3DPW train)	**51.7**	**85.7**	**99.5**	41.4	62.0	67.0	104.7

Table 2: **Evaluation of state-of-the-art models on 3DPW, Human3.6M, and MPI-INF-3DHP datasets.** DSR is our proposed model trained on monocular images similar to [14, 15, 18]. DSR outperforms all state-of-the-art models, including EFT [14] on the challenging datasets. "–" shows the results that are not available.

Method	PAMPJPE ↓	MPJPE ↓	PVE ↓
Standard Loss (SD)	47.5	73.9	99.2
SD + DSR	**45.1**	**71.3**	**96.6**

Table 3: **Potential of DSR.** We train and test on a subset of Human3.6M to evaluate the full potential of DSR loss. SD refers to standard joint loss.

Method	Ankle	Knee	Hip	Wrist	Elbow	Head
Standard Loss (SD)	99.3	54.6	20.0	109.5	81.1	81.8
SD + DSR	**96.1**	**50.4**	**19.5**	**107.3**	**79.3**	**80.5**

Table 4: **Per joint error for Human3.6M subset.** SD refers to standard joint loss used in 3D body estimation.

term. One possible explanation could be that the off-the-shelf segmentation model is not robust for hands and feet hence causing less gain. Empirically, we observe that soft-IoU performs better than soft-DistM and hence use it as the metric for DSR-MC for all subsequent experiments. Overall, the best accuracy is reached when both terms are used showing that supervising minimally-clothed and clothed regions differently helps improve 3D body estimation.

5.2. State-of-the-art comparison

We compare our approach with state-of-the-art methods in Table 2. We use two variants of our model, with and without the 3DPW training set, to be aligned with the training data of other methods. In 3DPW, an in-the-wild challenging 3D dataset, we outperform previous work when using 3DPW training data, while performing on par with EFT [14] when they are not used. Moreover, we clearly improve accuracy on Human3.6M [13], a standard indoor benchmark, over state-of-the-art SPIN [18] and EFT [14] methods. We also report on par results in MPI-INF-3DHP [2]. We perform significantly better than previous approaches that use ground-truth part-segmentation or silhouettes [30, 34, 52] compared to our weak supervision. Overall, we consistently perform better than previous approaches across different datasets, both indoors and outdoors. In Fig. 5 we can see different comparisons of DSR with the previous state-of-the-art and observe that the estimated mesh is more aligned with image evidence. These observations validate our hypothesis that clothing semantics, even when used as weak supervision, provides additional information for estimating more accurate 3D bodies.

5.3. Potential of DSR

To test the significance of high-level semantics on shape and pose estimation, we use an off-the-shelf segmentation model [8]. However, such models are not robust to in-the-wild examples. Because we use the output of the model as pseudo ground-truth for supervision, it is hard to determine the full potential of our approach. Hence, we experiment on the Human3.6M dataset to test the DSR loss in a more controlled setting. Human3.6M is an indoor dataset with significantly less background complexity as compared to outdoor datasets. Hence, it is ideal for testing the limit of DSR. We study two different cases. First, we split the training set of Human3.6M, with SMPL parameters computed by MoSh [25], into training and validation sets with S8 in the validation set. This is done to evaluate the per-vertex-error (PVE) using the MoSh ground truth SMPL parameters, thus, giving insight into the shape estimation efficacy of our method. As shown in Table. 3, the performance gain with the DSR loss is significantly higher compared to the standard joint loss. This emphasizes the importance of semantic information. We also analyze the per joint error to understand the source of a performance gain as shown in

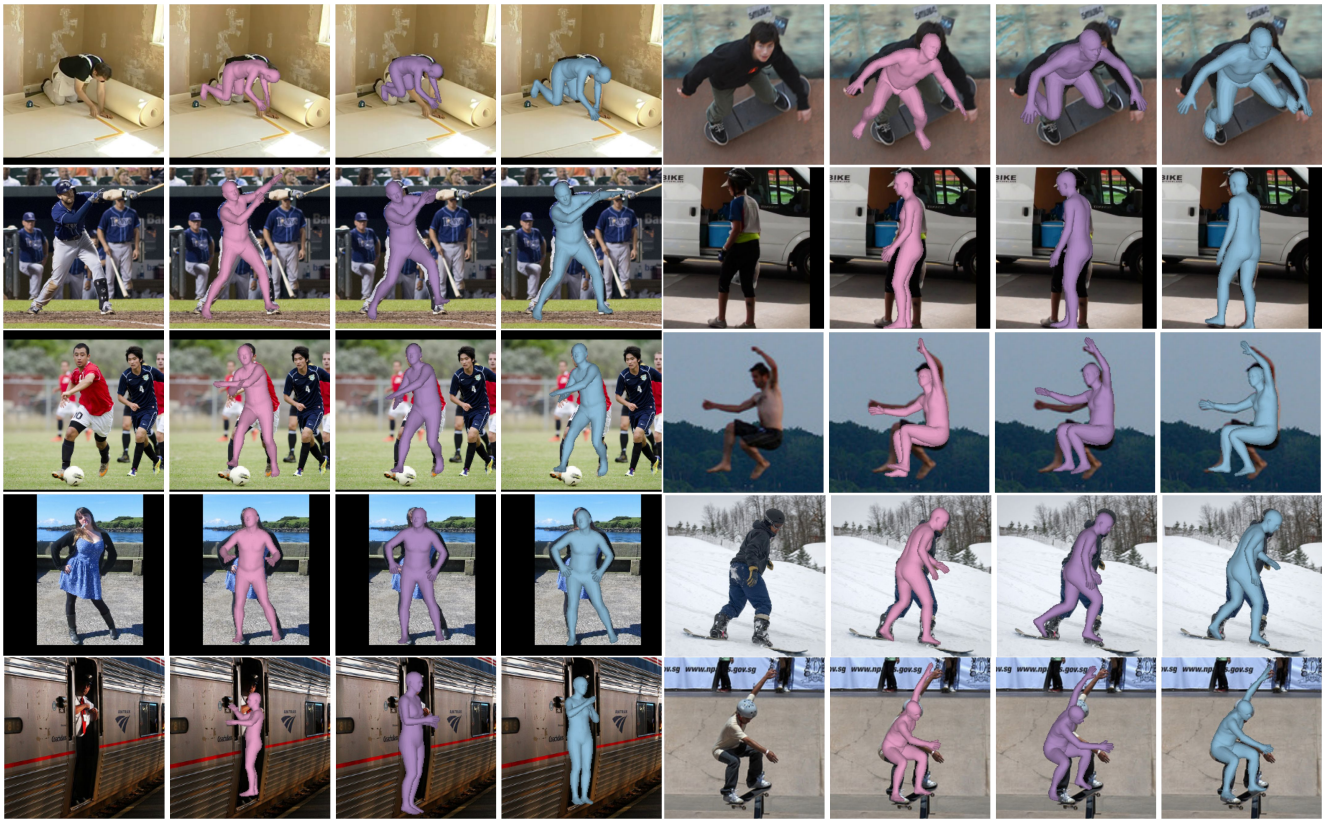

Figure 5: **Qualitative Results on COCO.** From left to right - Input image, SPIN [18], EFT [14] and DSR results.

the Table. 4. Using the DSR, the maximum performance gains are from *Ankle, Knee, Wrist* which are common failures in 3D pose estimation. Second, we take a step further to examine whether ground truth 3D joints are enough for accurate and pixel aligned body estimation. To this end, we take a random batch of 64 samples from Human3.6M and over-fit on only joints and joints with the DSR loss for 100 iterations with the same hyper-parameters used for other experiments. The qualitative results are depicted in Fig. 4. As we can see, supervision with ground 3D joints cannot always reason about all the pixels. Using DSR produces more pixel-aligned fits, especially for hands and feet.

6. Conclusion

While huge progress has been made in estimating 3D human pose and shape, we are still far from estimating highly accurate 3D humans in everyday scenes. We hypothesize that clothing semantics is an under-explored feature that can benefit 3D body estimation methods. Therefore, we introduce a novel method to exploit clothing semantics as weak supervision. Namely, we: (1) Introduce a novel differentiable loss that supervises clothed and minimally-clothed regions differently to ensure that the body lies inside the clothes for the former while tightly fitting for the latter. (2) Learn a semantic clothing prior, i.e. a probability distribution over clothing labels for SMPL vertices, to apply our method efficiently. This can also be used independently. (3) Thoroughly evaluate our approach qualitatively and quantitatively, outperforming the state-of-the-art. (4) Analyze our method's components and show that clothing semantics, even as weak supervision, is a valuable complementary cue to 3D joints that improves the estimation of 3D bodies. Our experiments show the importance of such semantics, providing new insight into 3D human body estimation.

DSR uses clothing as weak supervision, which can be limited in complex scenes with multiple people and occlusion. Our method can be easily extended to pipelines that account for multiple people in the scene [53]. In the future, we should explore methods that model 3D clothing semantics, build a better prior for SMPL bodies or incorporate additional constraints to disambiguate scene semantics.

Acknowledgements: We thank Sergey Prokudin, Chun-Hao P. Huang, Vassilis Choutas, Priyanka Patel, Radek Danecek, Cornelia Kohler and all Perceiving Systems department members for their help, feedback and fruitful discussions. **Disclosure:** https://files.is.tue.mpg.de/black/CoI/ICCV2021.txt

References

[1] Renderpeople. https://renderpeople.com, 2020. 4

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[3] Dragomir Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and James Davis. SCAPE: shape completion and animation of people. In *SIGGRAPH*, 2005. 3

[4] A. Balan and M. J. Black. The naked truth: Estimating body shape under clothing,. In *European Conference on Computer Vision (ECCV)*, 2008. 3, 5

[5] 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, 3

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 3

[7] Y. Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. In *Comput. Vis. Image Underst.*, 2020. 2

[8] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 4, 7

[9] P. Guan, A. Weiss, A. Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *International Conference on Computer Vision (ICCV)*, 2009. 3

[10] Riza Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[11] Tong He, J. Collomosse, H. Jin, and Stefano Soatto. Geopifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Advances in Neural Information Processing (NeurIPS)*, 2020. 3

[12] David C. Hogg. Model-based vision: a program to see a walking person. In *Image Vis. Comput.*, 1983. 3

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. In *IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 6, 7

[14] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *arXiv preprint arXiv:2004.03686*, 2020. 1, 3, 4, 5, 6, 7, 8

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 4, 6, 7

[16] A. Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[18] 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 *International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 6, 7, 8

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 7

[20] Christoph Lassner, J. Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and P. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3

[21] Sijin Li and Antoni B. Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*, 2014. 3, 6

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 3, 6

[23] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *International Conference on Computer Vision (ICCV)*, 2019. 4, 6

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 3

[25] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. In *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2014. 7

[26] Naureen Mahmood, N. Ghorbani, N. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[27] J. Martinez, Rayat Hossain, J. Romero, and J. Little. A simple yet effective baseline for 3D human pose estimation. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[28] Dushyant Mehta, H. Rhodin, D. Casas, P. Fua, Oleksandr Sotnychenko, Weipeng Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*, 2017. 6

[29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[30] 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 *International Conference on 3D Vision (3DV)*, 2018. 1, 2, 3, 7

[31] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4

[32] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[33] Georgios Pavlakos, Xiaowei Zhou, K. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[34] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 7

[35] U. Rafi, Juergen Gall, and B. Leibe. A semantic occlusion model for human pose estimation from a single depth image. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015. 3

[36] S. Saito, Zeng Huang, R. Natsume, S. Morishima, A. Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, 2019. 3

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

[38] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. In *Computer Vision and Image Understanding*, 2016. 2

[39] L. Sigal, S. Bhatia, S. Roth, Michael J. Black, and M. Isard. Tracking loose-limbed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004. 3

[40] Edgar Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A joint model for 2D and 3D pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[41] Carsten Stoll, N. Hasler, Juergen Gall, H. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *International Conference on Computer Vision (ICCV)*, 2011. 3

[42] K. Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[43] X. Sun, J. Shang, Shuang Liang, and Y. Wei. Compositional human pose regression. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[44] H. F. Tung, H. Tung, Ersin Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing (NeurIPS)*, 2017. 3

[45] G. Varol, D. Ceylan, Bryan C. Russell, Jimei Yang, Ersin Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[46] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 6

[47] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[48] Donglai Xiang, Fabián Prada, Chenglei Wu, and J. Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *International Conference on 3D Vision (3DV)*, 2020. 3

[49] Bin Xiao, Haiping Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[50] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3D pose and shape estimation by dense render-and-compare. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[51] S. Yu, Ye Yun, L. Wu, Gao Wenpeng, Fu Yi-li, and Mei Tao. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[52] A. Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, B. Freeman, R. Sukthankar, and C. Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 7

[53] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[54] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3D human mesh regression with dense correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[55] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3