PARE: Part Attention Regressor for 3D Human Body Estimation

Muhammed Kocabas[1,2] Chun-Hao P. Huang[1] Otmar Hilliges[2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zurich
{mkocabas,paul.huang,black}@tue.mpg.de otmar.hilliges@inf.ethz.ch

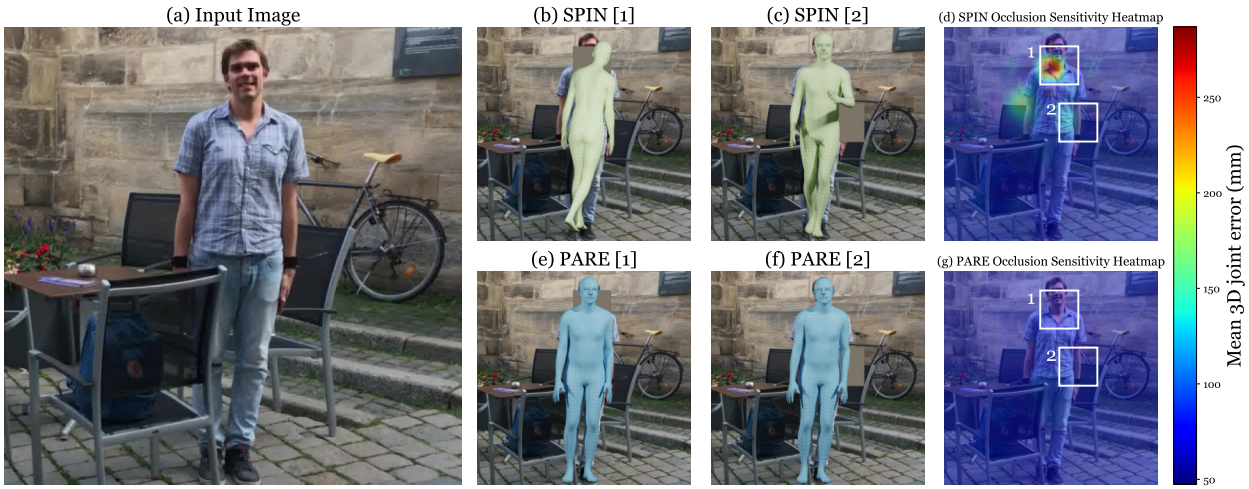

Figure 1: **Occlusion Sensitivity Analysis.** Given an input image (a), a small occluding patch (shown in gray) causes SPIN [29] to fail (b,c), whereas our method (PARE) (e,f) is robust to the occluder. Sub-figures on the right show the sensitivity of SPIN (d) and PARE (g) to an occluding patch (the size of the white squares) centered at every point in the image. Warmer colors mean higher average joint error.

Abstract

Despite significant progress, we show that state of the art 3D human pose and shape estimation methods remain sensitive to partial occlusion and can produce dramatically wrong predictions although much of the body is observable. To address this, we introduce a soft attention mechanism, called the Part Attention REgressor (PARE)*, that learns to predict body-part-guided attention masks. We observe that state-of-the-art methods rely on global feature representations, making them sensitive to even small occlusions. In contrast, PARE's part-guided attention mechanism overcomes these issues by exploiting information about the visibility of individual body parts while leveraging information from neighboring body-parts to predict occluded parts. We show qualitatively that PARE learns sensible attention masks, and quantitative evaluation confirms that PARE achieves more accurate and robust reconstruction results than existing approaches on both occlusion-specific and standard benchmarks. The code and data are available for research purposes at* https://pare.is.tue.mpg.de/

1. Introduction

Regressing 3D human pose and shape (HPS) directly from RGB images has many applications in robotics, computer graphics, AR/VR and beyond. The task is to take a single image [24, 29, 40] or video sequence [25, 27, 35] as input and to regress the parameters of a human body model such as SMPL [33] as output. Powered by deep CNNs, this task has seen rapid progress [24, 27, 29, 40]. However, in fully in-the-wild settings, people often appear under occlusion either due to self-overlapping body-parts, due to close-range interaction with other people or due to occluding objects such as furniture or other scene content. While pose estimation under occlusion has been treated in the literature [8, 9, 14, 19, 42, 43, 53, 54, 59], we highlight that this issue is particularly important in the context of direct regression methods. Such methods use all the pixels in the input to predict a single set of pose and shape parameters. Thus their pose estimates are particularly sensitive to even small perturbations in the observations of the body and its parts.

In this paper, we apply a visualization technique [58] for occlusion sensitivity analysis that yields insights into when

and why such methods fail. This indicates that, for state-of-the-art (SOTA) methods, relatively small occlusions, even of only a single joint, can lead to entirely implausible pose predictions. This is illustrated in Fig. 1, where we slide an occluder over the image, regress body pose, and compute the average 3D joint error with respect to ground truth. The heatmaps in Fig. 1 (d,g) illustrate a method's sensitivity to a square occluder centered at each pixel location (shown in white). The visualization reveals that methods like SPIN [40] are highly sensitive to localized part occlusion. To address this issue, we propose a method, based on a novel part-guided attention mechanism, making direct regression approaches more robust to occlusion.

The proposed method is called Part Attention REgressor (PARE). It has two tasks: the primary one is learning to regress 3D body parameters in an end-to-end fashion, and the auxiliary task is learning attention weights per body part. Each task has its own pixel-aligned feature extraction branch. We guide the attention branch with part segmentation labels in the early stages of training and continue without them for the later stages, thus we call it *body-part-driven attention*. Our key insight is that, to be robust to occlusions, the network should leverage pixel-aligned image features of visible parts to reason about occluded parts.

Given the success of attention-based methods on other tasks [11, 18, 34, 55], we exploit insights gained from the occlusion sensitivity analysis to focus attention on body parts. Therefore, we supervise the attention mask with part segmentations, but then train end-to-end with pose supervision only, allowing the attention mechanism to leverage all useful information from the body and the surrounding pixels. This gives the network freedom to attend to regions it finds informative in an unsupervised way. As a result, PARE learns to rely on visible parts of the body to improve robustness to occluded parts and overall performance on 3D pose estimation (Fig. 1 e-f).

To quantitatively evaluate the performance of PARE, we perform experiments on the 3DPW [52], 3DOH [59], and 3DPW-OCC [52] datasets. The results show that PARE yields consistently lower error than the state-of-the-art for both occlusion and non-occlusion cases.

In summary, our key contributions are: (1) We apply a visualization technique [58] to study how local part occlusion can influence global pose; we call this occlusion sensitivity analysis. (2) This analysis motivates a novel body-part-driven attention framework for 3D HPS regression that leverages pixel-aligned localized features to regress body pose and shape. (3) The network uses part visibility cues to reason about occluded joints by aggregating features from the attended regions, and by doing so, achieves robustness to occlusions. (4) We achieve SOTA results on a 3D pose estimation benchmark featuring occluded bodies, as well as a standard benchmark.

2. Related Work

We focus on 3D human shape and pose estimation from RGB images and discuss how previous approaches handle occlusions in various scenarios, e.g. self occlusion, camera frame occlusion, and scene object occlusion.

3D pose and shape from a single image. In estimating human shape and pose, many methods output the parameters of 3D human body models [3, 33, 39]. Initial work predicts the 3D body using keypoints and silhouettes [1, 4, 5, 15, 46]. These approaches are fragile, need manual input, use additional data, e.g. multi-view images, or do not generalize well to in-the-wild images. SMPLify [7] was the first automated method to fit the SMPL model to the output of a 2D keypoint detector [41]. Lassner et al. [31] employ silhouettes together with keypoints during fitting. In contrast, deep neural networks regress SMPL parameters directly from pixels [16, 24, 38, 40, 50, 51]. In order to deal with the lack of in-the-wild 3D ground-truth, methods use a 2D keypoint re-projection loss as weak supervision [24, 50, 51], use intermediate 2D representations, e.g. body/part segmentation [38, 40, 57], 2D sparse keypoints [45, 57], or leverage a human in the loop [31]. Note that the use of part segmentation in [31, 38, 57] is very different from our approach, in which part segmentations are used to facilitate soft attention. Kolotouros et al. [29] combine HMR [24] and SMPLify [7] in a training loop. At each step, HMR initializes SMPLify, which fits the body model to 2D joints, resulting in better supervision for the network. The above methods are typically sensitive to occlusion.

Implicit occlusion handling (data augmentation). Ideally, the regressed 3D body should be the same with or without occlusion. Current SOTA pose and shape estimation methods [24, 27, 29] directly encode the entire input region as one CNN feature after global average pooling, followed by body model parameter regression. The lack of pixel-aligned structure makes it hard for networks to explicitly reason about the locations and visibility of body parts. A common way to achieve robustness to occlusion in these frameworks is through data augmentation. For example, frame occlusion is often simulated by cropping [6, 23, 43], whereas object occlusion is approximated by overlaying object patches on the image [13, 44]. Instead of applying augmentation to input images, Cheng et al. [8] apply augmentations to heatmaps that contain richer semantic information and hence occlusions can be simulated in a more intelligent way. While helpful, these synthetic occlusions do not fully capture the complexity of occlusions in realistic images, nor do they provide insight into how to improve the network architecture to be inherently more robust to occlusion.

Explicit occlusion handling. To reason more explicitly about occlusions, previous work exploits visibility information. For example, Cheng et al. [9] avoid including occluded joints when computing losses during training. Such visi-

Figure 2: **Occlusion sensitivity analysis.** Heatmaps illustrate the error of SPIN [29] in individual joints caused by an occluder placed at each image location. Image size: 224×224; occluding patch: 40×40. The title of each heatmap names the joint and notes the range of the 3D error in mm visualized in the heatmap. See Section 3 for analysis.

bility information is obtained by approximating the human body as a set of cylinders, which is not realistic and only handles self occlusion. Wang et al. [54] learn to predict occlusion labels to zero out occluded keypoints before applying temporal convolution over a sequence of 2D keypoints.

Person-person occlusion is particularly common and challenging. For multi-person regression, Jiang et al. [21] use an interpenetration loss to avoid collision and an ordinal loss to resolve depth ambiguity. Sun et al. [56] estimate all people in an image simultaneously, enabling their method to learn about person-person occlusion. While [56] learns features that are robust to person-person occlusion, PARE learns to focus attention on individual body parts.

Zhang et al. [59] leverage saliency masks as visibility information to gain robustness to scene/object occlusions. Human meshes are parameterized by UV maps where each pixel stores the 3D location of a vertex, and occlusions are cast as an image-inpainting problem. The requirement of accurate saliency maps limits the performance on in-the-wild images. Furthermore, UV-coordinates can result in mesh artifacts, as shown in Sup. Mat.

3. Occlusion Sensitivity Analysis

To extract features from the input image region I, current direct regression approaches [24, 29] use a ResNet-50 [17] backbone and take the features after global average pooling (GAP), followed by an MLP that regresses and refines the parameters iteratively. In this section, we investigate the impact of occlusions on this type of architecture. Our analysis is inspired by Zeiler et al. [58] who systematically cover different portions of the image with a gray square to analyze how feature maps and classifier output changes. In contrast, we slide a gray occlusion patch over the image and regress body poses using SPIN [29]. Instead of computing a classification score as in [58], we measure the per-joint Euclidean distance between ground truth and predicted joints. We create an error heatmap, in which each pixel indicates how much error the model creates for joint j when the occluder is centered on this pixel. In addition to per-joint heatmaps, we compute an aggregate occlusion sensitivity map, that shows how the average joint error is influenced by an occlusion; this is visualized in Fig. 1(d) and in greater detail in the Sup. Mat.

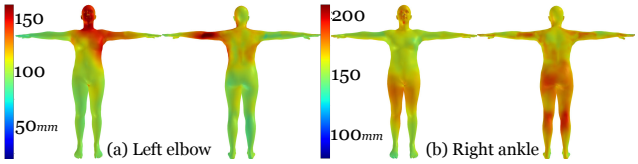

Figure 3: Occlusion sensitivity meshes for SPIN [29].

The per-joint error heatmaps for SPIN are visualized in Fig. 2 for a sample image from the 3DPW dataset [52]. Each sub-image corresponds to a particular joint and hot regions are locations where occlusion causes high error in this joint. This visualization allows us to make several observations. (1) Errors are low in the background and high on the body. This shows that SPIN has learned to attend to meaningful regions. (2) Joints visible in the original image have high errors when they are occluded by the square, as expected. (3) For joints that are naturally occluded, the network relies on other regions to reason about the occluded poses. For example, in the top row of Fig. 2, we observe high errors for the left/right ankles (which are occluded) when we occlude the thigh region. Since the network has no image features for the occluded parts, it must look elsewhere in the image for evidence. (4) Such dependencies happen not only between neighboring parts; occlusion can have long-range effects (e.g. occluding the pelvis causes errors in the head).

We further overlay the estimated body on the heatmap

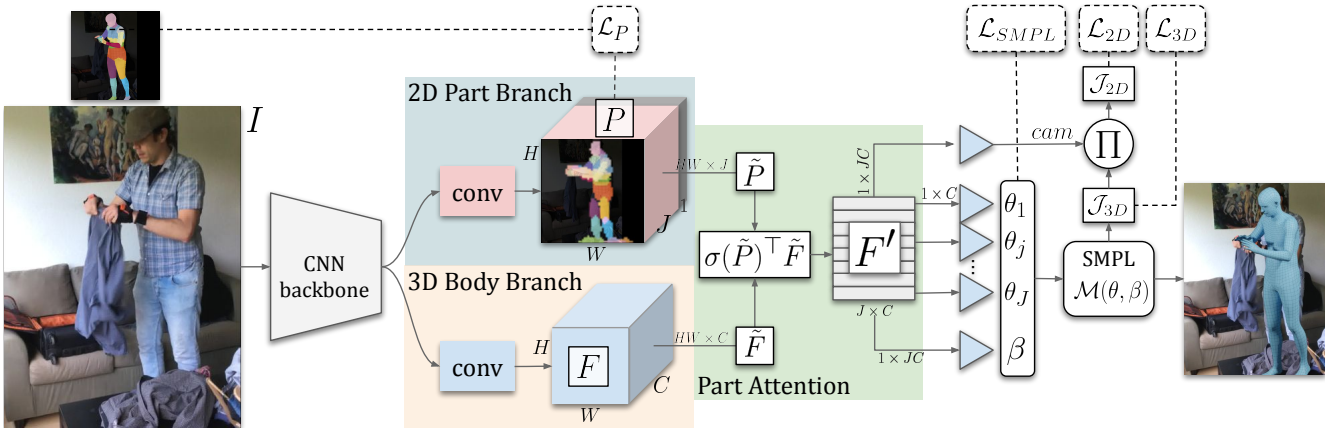

Figure 4: **PARE model architecture.** Given an input image, PARE extracts two pixel-level features P and F, which are fused by part attention (green box) leading to the final feature F' for camera and SMPL body regression.

to transfer the per-pixel error to visible vertices. We run this analysis over the complete 3DPW dataset, pool the per-vertex error across the dataset and visualize the result on a SMPL body model, giving one *occlusion sensitivity mesh* per joint. For example, Fig. 3(a) shows that the left elbow is sensitive to occlusion of the face, the left shoulder and the left upper arm region. See Sup. Mat. for more examples.

4. Method

Given the observations above, PARE is designed with the following insights. First, as shown in Fig. 2, SOTA networks [24, 27, 29] learn to attend to meaningful regions implicitly, despite limited spatial information after global average pooling. To better understand whether body parts are visible or not, and to know if their locations are occluded, PARE exploits a pixel-aligned structure, where each pixel corresponds to a region in the image and stores a pixel-level representation, namely, a feature volume. Second, since estimating attention weights and learning end-to-end trainable features for 3D poses are two different tasks, PARE is equipped with two feature volumes: one from the 2D part branch that estimates attention weights and one from the 3D body branch that performs SMPL parameter regression. Finally, to model the body part dependencies observed above, PARE exploits part segmentations as soft attention masks to adjust the contribution of each feature in the 3D body branch differently for each joint.

Preliminaries: Body Model. SMPL [33] represents the body pose and shape by Θ, which consists of the pose $\theta \in \mathbb{R}^{72}$ and shape $\beta \in \mathbb{R}^{10}$ parameters. Here we use the gender-neutral shape model as in previous work [24, 29]. Given these parameters, the SMPL model is a differentiable function that outputs a posed 3D mesh $\mathcal{M}(\theta, \beta) \in \mathbb{R}^{6890 \times 3}$. The 3D joint locations $\mathcal{J}_{3D} = W\mathcal{M} \in \mathbb{R}^{J \times 3}$, $J = 24$, are computed with a pretrained linear regressor W.

4.1. Model Architecture and Losses

The overall framework of PARE is depicted in Fig. 4. Our architecture works as follows: given an image I, we first run a CNN backbone to extract *volumetric* features, e.g. before the global average pooling layer for ResNet-50, followed by two separate feature extraction branches to obtain two volumetric image features. We denote the 2D part branch as $P \in \mathbb{R}^{H \times W \times (J+1)}$, modelling J part attention and 1 background masks, where H and W are the height and width of the feature volume and each pixel (h, w) stores the likelihood of belonging to a body part j. The other branch, denoted by $F \in \mathbb{R}^{H \times W \times C}$, is used for 3D body parameter estimation. It has the same spatial dimensions $H \times W$ as P but a different number of channels, C.

Let $P_j \in \mathbb{R}^{H \times W}$ and $F_c \in \mathbb{R}^{H \times W}$ denote the j-th and c-th channel of P and F, respectively, and let $F' \in \mathbb{R}^{J \times C}$ represent the final feature tensor. Each element in F_c contributes proportionally to F' according to the corresponding elements in P_j after spatial softmax normalization σ. Formally, the element at location (j, c) in F' is computed as:

$$F'_{j,c} = \sum_{h,w} \sigma(P_j) \odot F_c, \tag{1}$$

where $\odot$ is the Hadamard product. In other words, we use $\sigma(P_j)$ as a soft attention mask to aggregate features in F_c. This operation can be efficiently implemented as a dot product similar to existing attention implementations: $F' = \sigma(\tilde{P})^\top \tilde{F}$, where $\tilde{P} \in \mathbb{R}^{HW \times J}$ and $\tilde{F} \in \mathbb{R}^{HW \times C}$ denote the reshaped P (omitting the background mask) and F respectively. This attention operation suggests that if a particular pixel has a higher attention weight, its corresponding feature contributes more to the final representation F'. We supervise the 2D part branch P with ground-truth segmentation labels, which helps the attention maps of *visible* parts converge to the corresponding regions. For *occluded* parts, however, this encourages 0 attention weights for all pixels

in P_j because they do not exist in the ground-truth segmentation labels. An attention map with all 0 weights is undesirable and, in practice, also impossible since the spatial softmax ensures that all elements sum to 1. Therefore, we adopt a hybrid approach that supervises the 2D part branch only for the initial stage and continues training without any supervision. This allows the network to attend to other regions to estimate the poses of an occluded joint.

We take the full feature tensor F' to regress body shape β and a weak-perspective camera model with scale and translation parameters $[s, t], t \in \mathbb{R}^2$, while each row, F'_j, is also sent to different MLPs to predict the rotation of each part, θ_j, parameterized as a 6D vector following [27, 29] [1].

Overall, our total loss is:

$$\mathcal{L} = \lambda_{3D}\mathcal{L}_{3D} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{SMPL}\mathcal{L}_{SMPL} + \lambda_P\mathcal{L}_P, \quad (2)$$

where each term is calculated as:

$$\begin{aligned}
\mathcal{L}_{3D} &= \|\mathcal{J}_{3D} - \hat{\mathcal{J}}_{3D}\|_F^2, \\
\mathcal{L}_{2D} &= \|\mathcal{J}_{2D} - \hat{\mathcal{J}}_{2D}\|_F^2, \\
\mathcal{L}_{SMPL} &= \|\Theta - \hat{\Theta}\|_2^2, \\
\mathcal{L}_P &= \frac{1}{HW}\sum_{h,w} \text{CrossEntropy}\left(\sigma(P_{h,w}), \hat{P}_{h,w}\right),
\end{aligned}$$

where $\hat{x}$ represents the ground truth for the corresponding variable x. To compute the 2D keypoint loss, we need the SMPL 3D joint locations $\mathcal{J}_{3D}(\theta, \beta) = W\mathcal{M}(\theta, \beta)$, which are computed from the body vertices with a pretrained linear regressor W. With the inferred weak-perspective camera, we compute the 2D projection of the 3D joints $\mathcal{J}_{3D}$, as $\mathcal{J}_{2D} \in \mathbb{R}^{J\times 2} = s\Pi(R\mathcal{J}_{3D}) + t$, where $R \in SO(3)$ is the camera rotation matrix and Π is the orthographic projection. λ is a scalar coefficient to balance the loss terms. Let $P_{h,w} \in \mathbb{R}^{1\times 1\times (J+1)}$ denote the fiber of P at the location (h, w), and $\hat{P}_{h,w} \in \{0, 1\}^{(J+1)}$ denotes the ground-truth part label at the same location, expressed as a one-hot vector. The part segmentation loss $\mathcal{L}_P$ is the cross-entropy loss between $P_{h,w}$ after softmax and $\hat{P}_{h,w}$, averaged over $H\times W$ elements. Note that this softmax normalizes along the fiber $P_{h,w}$ while the one in Eq. 1 normalizes over the slice P_j.

4.2. Implementation Details

As mentioned above, the body-part label supervision via $\mathcal{L}_P$ is applied on the attention tensor P only in the *initial* stages of training. It is later removed by setting λ_P to zero, turning the attention mechanism into an unsupervised pure soft-attention. The absence of body-parts due to occlusion is the main motivation for this training scheme. Setting λ_P to zero allows the attention mechanism to also consider pixels beyond the body itself. Hence, the final attention maps do not necessarily (and often do not) resemble body part segmentations, as shown later in Fig. 7 and Sup. Mat. If a body part is visible, it focuses on that part directly; if it is occluded, the attention is free to leverage other informative regions in the image. In Sec. 5, we analyze how the accuracy of part segmentation impacts body reconstruction.

[1] With slight abuse of notations, θ is in axis-angle form when passed to the SMPL model but in 6D-vector form during the regression and loss computation.

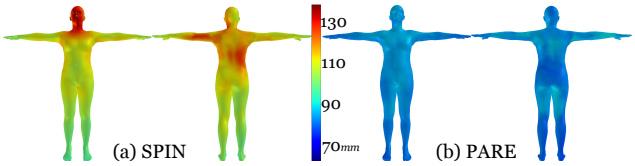

Figure 5: **Occlusion sensitivity mesh.** Meshes visualize the (a) SPIN and (b) PARE average joint errors.

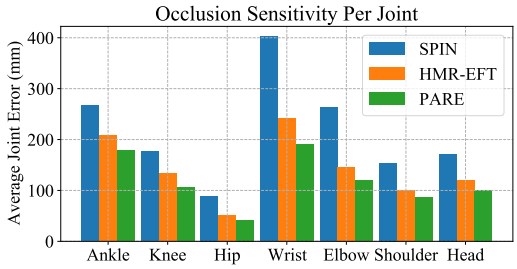

Figure 6: Per joint occlusion sensitivity analysis of three different methods: SPIN [29], HMR-EFT [23] (trained with occlusion augmentation), and PARE. PARE is consistently more robust to occlusion.

We evaluate both ResNet-50 [17] and HRNet-W32 [48] networks as the backbone. Since ResNet-50 is widely used in other SOTA methods [24, 27, 29], we choose it as the default backbone for most of the experiments unless stated otherwise. We extract the $7 \times 7 \times 2048$ feature volumes before global average pooling. For the 2D and 3D branches, we use three $2\times$ upsampling followed by 3×3 convolutional layers applied with batch-norm and ReLU. The number of conv kernels is 256. For HRNet-W32, since it already provides volumetric features with a higher resolution, we only use two 3×3 convolutional layers applied with batch-norm and ReLU as the 2D and 3D branches.

To obtain part attention maps, we apply $J + 1$ 1×1 convolutional kernels to 2D part features to reduce the channel dimension. After obtaining the $J \times C$ final feature F', we use separate linear layers to predict each SMPL joint rotation θ_j. We regress shape and camera parameters from the flattened F' vector. We use a fixed image size of 224×224 for all experiments. The Adam optimizer with a learning rate of 5×10^{-5} and batch size 64 is used to optimize our model. PARE is end-to-end trainable in a single stage, unlike recent multi-stage methods [10, 16, 37, 57].

	Method	3DPW MPJPE ↓	3DPW PA-MPJPE ↓	3DPW PVE ↓
temporal	HMMR [24]	116.5	72.6	-
	Doersch et al. [12]	-	74.7	-
	Sun et al. [49]	-	69.5	-
	VIBE [27]	93.5	56.5	113.4
	MEVA [35]	86.9	54.7	-
multi stage	Pose2Mesh [10]	89.2	58.9	-
	Zanfir et al. [57]	90.0	57.1	-
	I2L-MeshNet [37]	93.2	58.6	-
	LearnedGD [47]	-	56.4	-
single stage	HMR [24]	130.0	76.7	
	CMR [30]	-	70.2	-
	SPIN [29]	96.9	59.2	135.1
	HMR-EFT [23]	-	54.2	-
	PARE (R50)	82.9	52.3	99.7
	PARE (HRNet-W32)	82.0	50.9	97.9
	PARE (HRNet-W32) w. 3DPW	**74.5**	**46.5**	**88.6**

Table 1: **Evaluation on the `3DPW` dataset.** The units for mean joint and vertex errors are in *mm. PARE models outperform temporal, multi-stage, and single-stage state-of-the-art methods.*

Method	3DPW-OCC MPJPE ↓	3DPW-OCC PA-MPJPE ↓	3DPW-OCC PVE ↓	3DOH MPJPE ↓	3DOH PA-MPJPE ↓
Zhang et al. [59]	-	72.2	-	-	58.5
SPIN [29]	95.6	60.8	121.6	104.3	68.3
HMR-EFT [23]	94.4	60.9	111.3	75.2	53.1
PARE (R50)	**90.5**	**56.6**	**107.9**	**63.3**	**44.3**

Table 2: **Evaluation on occlusion datasets `3DPW-OCC`, `3DOH`.** Here all methods except SPIN are trained with the same datasets, i.e. `COCO`, `Human3.6M` and `3DOH`.

5. Experiments

Training. We train PARE on `COCO` [32], `MPII` [2], `LSPET` [22], `MPI-INF-3DHP` [36], and `Human3.6M` [20] datasets. More details about these datasets are provided in Sup. Mat. Pseudo-ground-truth SMPL annotations for in-the-wild datasets are provided by EFT [23]. The part segmentation labels are obtained through rendering segmented SMPL meshes, as visualized in Fig. 4. We use 24 parts corresponding to 24 SMPL joints. See Sup. Mat. for samples of part segmentation labels. We used the PyTorch reimplementation [28] of Neural Mesh Renderer [26] to render the parts. For samples without a part segmentation label, we do not supervise the 2D branch.

For the ablation experiments, we train PARE and our baselines on `COCO` for 175K steps and evaluate on `3DPW` and `3DPW-OCC` datasets. We then incorporate all the training data to compare PARE to previous SOTA methods. This pretraining strategy accelerates convergence and reduces the overall training time. It takes about 72 hours to train PARE until convergence on an Nvidia RTX2080Ti GPU.

To increase robustness to occlusion, we use common occlusion augmentation techniques; i.e. synthetic occlusion (SynthOcc) [44] and random crop (RandCrop) [23, 43]. All PARE and baseline HMR-EFT models are trained with SynthOcc augmentation unless stated otherwise, e.g. Table 4.

Evaluation. The `3DPW` [52] test split, `3DPW-OCC` [52, 59], and `3DOH` [59] datasets are used for evaluation. We report Procrustes-aligned mean per joint position error (PA-MPJPE) and mean per joint position error (MPJPE) in mm. For `3DPW` we also report per vertex error (PVE) in mm.

Comparison to the state-of-the-art. Table 1 compares PARE with previous single-RGB-image HPS estimation methods. We report PARE results with two different backbones: ResNet-50 and HRNet-W32. PARE improves the PA-MPJPE performance by 10% compared to HMR-EFT [23], one of the best-performing recent methods.

Table 2 demonstrates the performance of PARE on occlusion-specific datasets. Here Zhang et al. [59], HMR-EFT [23], and PARE are trained with `COCO`, `Human3.6M`, and `3DOH` for a fair comparison. We report the SPIN results for reference. HMR-EFT is the fair alternative to SPIN, since SPIN uses HMR as the architecture. PARE consistently improves the performance on these occlusion datasets. Although HMR-EFT is trained with exactly the same augmentation and data as PARE, it performs worse.

We also quantify our occlusion sensitivity analysis. Figure 5 shows the average joint error of SPIN and PARE methods on the `3DPW` test split. SPIN is quite sensitive to upper body occlusions, especially around the head and back. PARE is more robust to occlusions and yields lower error overall. See Sup. Mat. for the per-joint version of Fig. 5. Figure 6 shows the per-joint breakdown of the mean 3D error from the occlusion sensitivity analysis for three different methods, SPIN, HMR-EFT, and PARE. Here, we retrain HMR-EFT using SynthOcc for a fair comparison. Again, PARE improves the occlusion robustness of all joints.

Qualitative comparison. We qualitatively compare SPIN, HMR-EFT, and PARE in Fig. 8. Even though occlusion augmentation improves robustness to occlusion as seen in the HMR-EFT results, it is not sufficient on its own. PARE, with its attention mechanism, performs well even in challenging occlusion scenarios. More qualitative samples, including failure cases, are provided in Sup. Mat.

Does part attention help? Table 3 summarizes our ablation experiments that explore the concept of part attention. First, we compare our results with Neural Body Fitting [38] trained with identical settings to ours. NBF [38] can be seen as a straightforward combination of part segmentation and human body regression. Table 3 shows that NBF's two-stage approach is outperformed even by the HMR-EFT baseline. Subsequently, we compare different types of supervision for the 2D part branch P and sampling methods to obtain final features F' from F. "Unsup" means P is not supervised. Inspired by HoloPose [16], we first supervise the 2D branch with keypoints and pool the 3D features via bilin-

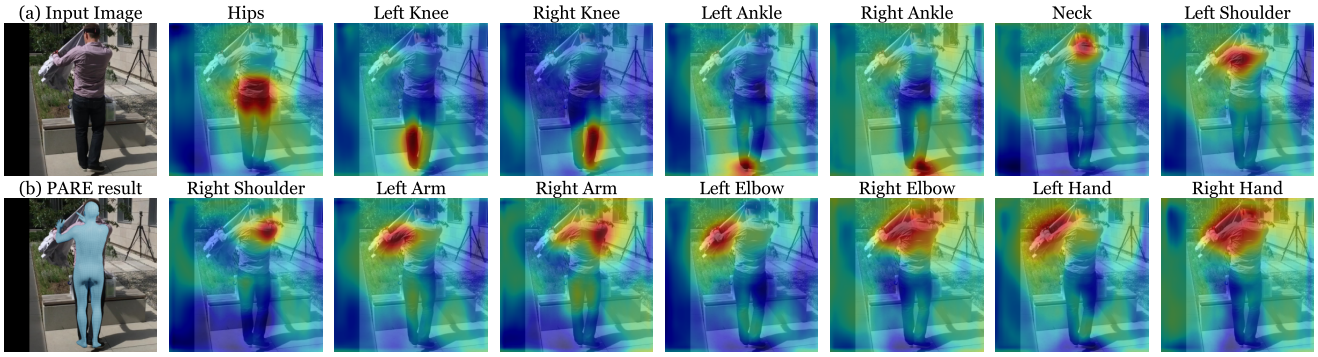

Figure 7: **PARE attention visualization.** Attention maps predicted by the 2D part branch for different joints in image (a). For occluded joints like row 2 right hand, PARE learns to attend to larger, more distant, regions to glean information.

		3DPW		3DPW-OCC	
Method		MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
	NBF [38]	100.4	63.2	103.5	70.4
	HMR-EFT	99.0	59.9	97.9	64.7
P **Supervision**	F **Sampling**				
(a) Joints	Pooling	95.2	58.9	95.4	63.1
(b) Joints	Attention	95.3	58.8	98.9	63.9
(c) Unsup	Attention	94.8	57.9	95.9	62.7
(d) Parts	Attention	94.5	57.3	94.7	**61.2**
(e) Parts/Unsup	Attention	**93.4**	**57.1**	**93.9**	61.6
(f) Parts	Pooling	97.9	59.1	99.8	64.8

Table 3: **Exploring part attention.** The "P Supervision" column shows the type of supervision for the 2D part branch P. "F Sampling" shows the type of feature sampling method for F. All methods are trained on COCO-EFT with a ResNet-50 backbone.

ear sampling (Table 3-a). Even though this gives lower error than HMR, the improvement is not significant. Intuitively, sparse keypoints do not cover enough spatial area to be able to reason about body parts. Because the 2D branch predicts Gaussian heatmaps, which cover a larger spatial area than discrete keypoints, we explore soft attention instead of pooling to have a larger effective receptive field (Table 3-b). In doing so, however, we do not leverage the full potential of soft attention, which can learn which regions to attend to implicitly from the data. So, we remove supervision for the 2D branch to see if soft attention alone can work as well as explicit supervision (Table 3-c). Upon visualizing the resulting attention maps, we find that they are not focused on the body parts. To induce more structure, we supervise the 2D branch with part segmentation labels (Table 3-d). This approach works significantly better than the above attempts. There is a remaining caveat, however: by supervising with a segmentation loss, we constrain the attention map to the parts only, whereas a pure soft attention has the potential to attend to any region it finds informative. Consequently, we train with mixed supervision, applying the part segmentation loss for around 125K steps, then continuing to train without supervision (Table 3-e). This final version produces

	3DPW		3DPW-OCC	
Method	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
HMR-EFT + SynthOcc	99.0	59.9	97.9	64.7
PARE	95.0	57.6	**94.4**	61.3
PARE + SynthOcc	**94.5**	**57.3**	94.7	**61.2**
PARE + SynthOcc + RandCrop	95.7	58.1	97.8	62.6

Table 4: **Ablation of different occlusion augmentation strategies.** We demonstrate the effect of synthetic occlusion (SynthOcc) and random crop (RandCrop) augmentation on the final performance. All methods are trained on COCO-EFT with ResNet-50 as the backbone.

		3DPW		3DPW-OCC	
Method		MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
HMR-EFT	ResNet-50	99.0	59.9	97.9	64.7
PARE	ResNet-50	**93.4**	**57.1**	**93.9**	**61.6**
HMR-EFT	HRNet-W32	92.6	55.9	90.2	57.8
PARE	HRNet-W32	**89.0**	**54.3**	**87.1**	**57.0**

Table 5: **Ablation of backbone architectures**. All methods are trained on COCO-EFT.

the "best of both worlds" and the lowest error. We also experiment with part segmentation and pooling to explore the effect of soft-attention (Table 3-f). Finally, to demonstrate the statistical significance, we performed a two-sided t-test for all experiments in Table 3; specifically $p<0.01$ for rows (c) vs. (d), (d) vs. (e), and (b) vs. (d).

In addition to joint errors, we measure the mean part segmentation IoU (intersection over union) to better understand how part segmentation and the final pose and shape estimation interact when we do not use part supervision. Mean IoU on the 3DPW test set is 1%, 85%, 74% for (c) unsup, (d) parts, and (e) parts/unsup methods respectively. Lower segmentation accuracy does not hurt the body reconstruction. We provide further body-part segmentation results during different stages of the training in Sup. Mat.

Figure 7 visualizes these attention maps on sample images. Part attention learns to attend to body parts or image regions as needed to estimate body shape and pose.

Occlusion Augmentation. We report the effect of occlu-

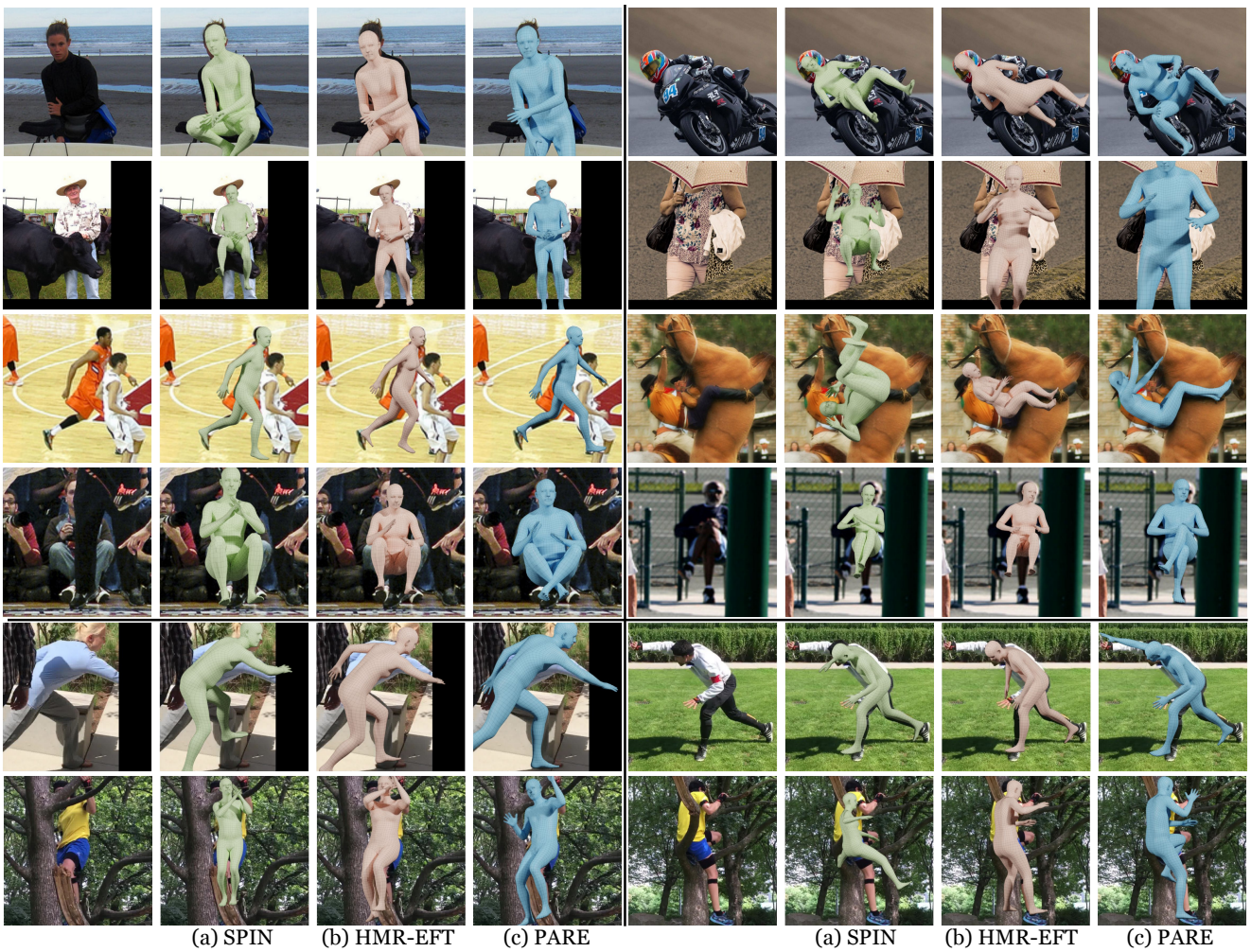

Figure 8: **Qualitative results on COCO (rows 1-4) and 3DPW (rows 5-6) datasets.** From left to right: Input image, (a) SPIN [29] results, (b) HMR-EFT [23] results, (c) PARE results.

sion augmentation techniques in Table 4. SynthOcc improves the performance on both 3DPW and 3DPW-OCC over vanilla training. Applying RandCrop right at the beginning of the training hurts the performance. Therefore, we start applying crop augmentation after 175K training steps. Between 30%-50% of a bounding box is cropped with the probability of 0.3. Even though crop augmentation does not improve performance on 3DPW and 3DPW-OCC, we find it useful for true in-the-wild images, which often contain significant frame occlusion. See Sup. Mat. for more examples.

Effect of CNN backbones. As shown in Table 5, HRNet-W32, which produces effective high-resolution representations, performs better than ResNet-50. PARE provides consistent improvements over HMR-EFT with both backbones.

6. Conclusion

We present a novel Part Attention Regressor, PARE, which regresses 3D human pose and shape by exploiting information about the visibility of individual body parts, and thus gaining robustness to occlusion. PARE is based on the insights gleaned from our occlusion sensitivity analysis. In particular, we observe dependencies between body parts and argue that the network should rely on visible parts to improve predictions for occluded parts and, hence, the overall performance of 3D pose estimation. Our novel body-part-driven attention mechanism captures such dependencies, using soft attention guided by regressed body part segmentation masks. The network learns to use part segmentations as visibility cues to reason about occluded joints and aggregating features from the attended regions. This improves robustness to occlusions of different types: scene, self, and frame occlusion. Detailed ablation studies show how each choice contributes to our state-of-the-art performance on benchmark datasets.

References

[1] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006. 2

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 6

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *SIGGRAPH*, 2005. 2

[4] Alexandru Balan and Michael J Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, 2008. 2

[5] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007. 2

[6] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. In *Advances in Neural Information Processing*, 2020. 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*, 2016. 2

[8] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *arXiv preprint arXiv:2004.11822*, 2020. 1, 2

[9] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *International Conference on Computer Vision*, pages 723–732, 2019. 1, 2

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787, 2020. 5, 6

[11] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40, 2020. 2

[12] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3D pose estimation: motion to the rescue. In *Advances in Neural Information Processing*, 2019. 6

[13] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision*, 2020. 2

[14] Golnaz Ghiasi, Yi Yang, Deva Ramanan, and Charless C Fowlkes. Parsing occluded people. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[15] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3D structure with a statistical image-based shape model. In *International Conference on Computer Vision*, pages 641–648, 2003. 2

[16] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 2, 5, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016. 3, 5

[18] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7779–7788, 2020. 2

[19] Jia-Bin Huang and Ming-Hsuan Yang. Estimating human pose from occluded images. In *Asian Conference on Computer Vision*, pages 48–60, 2009. 1

[20] 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*, 2014. 6

[21] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[22] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 6

[23] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. *arXiv:2004.03686*, 2020. 2, 5, 6, 8

[24] 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*, pages 7122–7131, 2018. 1, 2, 3, 4, 5, 6

[25] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 1

[26] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6

[27] 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*, pages 5252–5262, 2020. 1, 2, 4, 5, 6

[28] Nikos Kolotouros. Pytorch implementation of the neural mesh renderer, 2018. 6

[29] 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*, pages 2252–2261, 2019. 1, 2, 3, 4, 5, 6, 8

[30] 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*, pages 4501–4510, 2019. 6

[31] 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 Conference on Computer Vision and Pattern Recognition*, pages 4704–4713, 2017. 2

[32] 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*, 2014. 6

[33] 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. 1, 2, 4

[34] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[35] Zhengyi Luo, S. Golestaneh, and Kris M. Kitani. 3D human motion estimation via motion compression and refinement. In *Asian Conference on Computer Vision*, pages 324–340, 2020. 1, 6

[36] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *International Conference on 3DVision*, 2017. 6

[37] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pages 752–768, 2020. 5, 6

[38] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *International Conference on 3DVision*, 2018. 2, 6, 7

[39] 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 Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2

[40] 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*, pages 459–468, 2018. 1, 2

[41] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[42] Umer Rafi, Juergen Gall, and Bastian Leibe. A semantic occlusion model for human pose estimation from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015. 1

[43] Chris Rockwell and David F. Fouhey. Full-body awareness from partial observations. In *European Conference on Computer Vision*, pages 522–539, 2020. 1, 2, 6

[44] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. How robust is 3d human pose estimation to occlusion? *IROS workshops*, 2018. 2, 6

[45] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *British Machine Vision Conference*, 2020. 2

[46] Leonid Sigal, Alexandru Balan, and Michael J Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing*, 2008. 2

[47] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, 2020. 6

[48] Ke 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*, 2019. 5

[49] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *International Conference on Computer Vision*, 2019. 6

[50] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human shape and pose prediction. In *British Machine Vision Conference*, 2017. 2

[51] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing*, pages 5236–5246, 2017. 2

[52] 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*, pages 614–631, 2018. 2, 3, 6

[53] Saeid Vosoughi and Maria A Amer. Deep 3d human pose estimation under partial body presence. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 569–573. IEEE, 2018. 1

[54] Justin Wang, Edward Xu, Kangrui Xue, and Lukasz Kidzinski. 3D pose detection in videos: Focusing on occlusion. *arXiv preprint arXiv:2006.13517*, 2020. 1, 3

[55] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[56] Sun Yu, Bao Qian, Liu Wu, Fu Yili, and Mei Tao. CenterHMR: a bottom-up single-shot method for multi-person 3d mesh recovery from a single image. 2020. 3

[57] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision*, pages 465–481, 2020. 2, 5, 6

[58] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014. 1, 2, 3

[59] 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*, pages 7374–7383, 2020. 1, 2, 3, 6