Learning to Estimate Robust 3D Human Mesh from In-the-Wild Crowded Scenes

Hongsuk Choi[1] Gyeongsik Moon[1] JoonKyu Park[1] Kyoung Mu Lee[1,2]

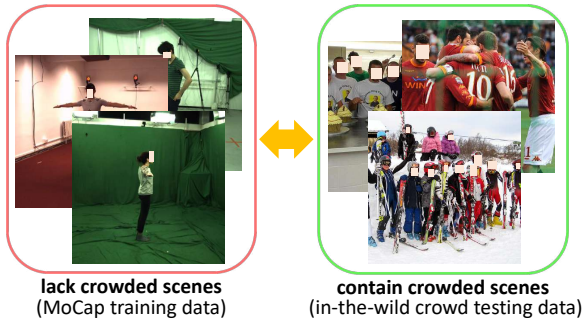
[1]Dept. of ECE & ASRI, [2]IPAI, Seoul National University, Korea
{redarknight,mks0601,jkpark0825,kyoungmu}@snu.ac.kr

Abstract

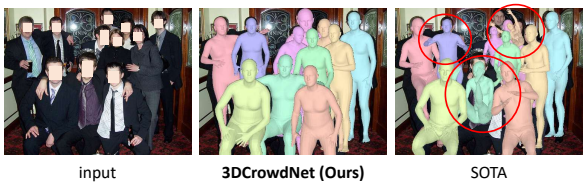
We consider the problem of recovering a single person's 3D human mesh from in-the-wild crowded scenes. While much progress has been in 3D human mesh estimation, existing methods struggle when test input has crowded scenes. The first reason for the failure is a domain gap between training and testing data. A motion capture dataset, which provides accurate 3D labels for training, lacks crowd data and impedes a network from learning crowded scene-robust image features of a target person. The second reason is a feature processing that spatially averages the feature map of a localized bounding box containing multiple people. Averaging the whole feature map makes a target person's feature indistinguishable from others. We present 3DCrowdNet that firstly explicitly targets in-the-wild crowded scenes and estimates a robust 3D human mesh by addressing the above issues. First, we leverage 2D human pose estimation that does not require a motion capture dataset with 3D labels for training and does not suffer from the domain gap. Second, we propose a joint-based regressor that distinguishes a target person's feature from others. Our joint-based regressor preserves the spatial activation of a target by sampling features from the target's joint locations and regresses human model parameters. As a result, 3DCrowdNet learns target-focused features and effectively excludes the irrelevant features of nearby persons. We conduct experiments on various benchmarks and prove the robustness of 3DCrowdNet to the in-the-wild crowded scenes both quantitatively and qualitatively. Codes are available here [1].

1. Introduction

Extensive research has been committed to reconstructing an accurate 3D human mesh, which represent both the pose and shape of a human, from a single image. However, 3D human mesh estimation from in-the-wild crowded scenes has been barely studied, despite their common presence. Consequently, most of the previous works show results on scenes without inter-person occlusion and provide inaccurate results on crowded scenes. The inter-person occlusion is the essential challenge of in-the-wild crowded scenes, and many practical applications including abnormal behavior detection [8] and person re-identification [35] encounter such situations. This paper investigates the limitation of the current literature and proposes a novel method for robust 3D human mesh estimation from in-the-wild crowded scenes.



(a) Domain gap (occlusion/pose/appearance, etc)



(b) Qualitative comparison

Figure 1. 3DCrowdNet resolves (a) a domain gap issue in estimating a 3D human mesh from in-the-wild crowded scenes. Due to the large domain gap between motion capture data and in-the-wild crowd data, (b) existing state-of-the-art methods such as SPIN [19] produce inaccurate results, while 3DCrowdNet gives an accurate 3D human mesh despite severe inter-person occlusion. We conceal a person's face in this paper to abide by the ethical policy.

[1] https://github.com/hongsukchoi/3DCrowdNet_RELEASE

The currently dominant training strategy for human mesh recovery is mixed-batch training. It composes a mini-batch with one-half data from a motion capture (MoCap) 3D dataset [13, 26] and the other from an in-the-wild 2D dataset [22]. To use the 2D dataset for supervision, 3D joints regressed from a predicted mesh are projected onto the image plane, and the distance with 2D annotations is computed. This way of mixing 3D and 2D data is well known to improve accuracy and generalization [17, 19] by implicitly inducing a neural network to benefit from accurate 3D annotations of the 3D data and diverse image appearances in the 2D data. The dominant approach of recent works [5, 9, 19] is a model-based approach using a global feature vector, which obtains the feature vector with a deep convolutional neural network (CNN) and regresses the human model parameters (*e.g.* SMPL [24]) from it. First, they crop an image using a bounding box of a target person detected from off-the-shelf human detectors [10]. Then they process the target's cropped image with a deep CNN and perform a global average pooling to obtain the global feature vector. The global feature vector is fed to a Multi-Layer Perceptron (MLP)-based regressor that regresses the mesh parameters. The 3D meshes are obtained by forwarding the parameters to the human model layers.

While the recent works have shown reasonable results on standard benchmarks [13,46] based on the two wheels of the current literature, in-the-wild crowded scenes remain insurmountable due to the following two reasons. First, a large domain gap between training data from MoCap datasets and testing data from in-the-wild crowded scenes hinders a deep CNN from extracting proper image features of a target person. The domain gap arises from the presence of a human crowd, which entails diverse inter-person occlusion, interacting body poses, and indistinguishable cloth appearances (Figure 1a). The mixed-batch training alone is insufficient to overcome the domain gap, and existing methods struggle to acquire robust image features from in-the-wild crowded scenes, and produce inaccurate meshes (Figure 1b). Intuitively, this tells us that external guidance robust to the domain gap is required for a crowded scene-robust image feature, in addition to the mixed-batch training.

Next, the global average pooling on a deep CNN feature collapses the spatial information that distinguishes a target person's feature from others. In-the-wild crowded scenes often involve overlapping people and inaccurate human bounding boxes. Thus, a bounding box of a target inevitably includes non-target people. A deep CNN feature retains features of these non-target people, and the global average pooling makes a target person's feature indistinguishable from others. This confuses a regressor and makes it difficult to capture an accurate 3D pose of a target person. For instance, the regressor may miss human parts occluded by another person or predict a different person's pose.

In this regard, we present 3DCrowdNet, a novel network that learns to estimate a single person's robust 3D human mesh from in-the-wild crowded scenes. This study is one of the earliest works that explicitly tackle 3D human mesh estimation of a target person in a crowd. 3DCrowdNet addresses the two issues of previous works in two folds. First, we resolve the domain gap by explicitly guiding a deep CNN to extract a crowded scene-robust image feature using an off-the-shelf 2D pose estimator. Unlike methods targeting 3D geometry, the 2D pose estimator does not require depth supervision and is not trained on a MoCap dataset. Instead, it is trained only on in-the-wild datasets [21, 41] that have images containing human crowds and suffers less from a domain gap regarding the inference on crowded scenes. Consequently, the 2D pose estimator's outputs provide strong evidence of a target person and help 3DCrowdNet pay attention to a target's feature despite the challenges in in-the-wild crowded scenes.

Second, we propose a joint-based regressor that does not blow away the spatial activation of a target person in a feature map with the global average pooling. The joint-based regressor first predicts the spatial locations of joints. Then, it samples image features from a deep CNN feature map with the locations. In particular, we keep the sampling area small to exclude features of non-target people. The target person's feature is distinguished from others, and human model parameters are regressed from the sampled image features. The joint-based regressor differs from the previous regressors that evenly aggregate people's features regardless of the target. Figure 2 depicts the overview of 3DCrowdNet.

Note that 3DCrowdNet substantially differs from prior works [6, 25] that directly lift 2D estimation outputs to 3D-(a) we focus on producing and leveraging image features of a target person in human crowds, and (b) such image features help 3DCrowdNet to resolve the depth and shape ambiguity of a target person, from which the 2D estimation outputs inherently suffer. Thus, we argue that this work takes a step towards accurate 3D human mesh estimation from in-the-wild crowded scenes by distinguishing image features of a target person in densely interacting crowds, which is highly challenging but important. The experiments show that 3DCrowdNet significantly outperforms the previous 3D human mesh estimation methods on in-the-wild crowded scenes. Also, it achieves state-of-the-art accuracy in multiple 3D benchmarks [16, 28, 46]. Extensive qualitative results are presented in the main manuscript and supplementary material. Our contributions can be summarized as follows:

- We present 3DCrowdNet, the first approach to 3D human mesh recovery from in-the-wild crowded scenes. It effectively processes image features of a target person in a crowd, which is essential for accurate 3D pose and shape reconstruction.

- It extracts crowded scene-robust image features by resolving the domain gap with a 2D pose estimator.
- It distinguishes a target person's image features from others using a joint-based regressor.
- 3DCrowdNet significantly outperforms previous methods on in-the-wild crowded scenes both quantitatively and qualitatively, and achieves state-of-the-art 3D pose and shape accuracy on multiple 3D benchmarks.

2. Related works

2D human pose estimation from crowded scenes. Early works of 2D human pose estimation did not explicitly target crowded scenes. However, their methods are related to diverse challenges of in-the-wild crowded scenes, such as overlapping human bounding boxes, human detection error, and inter-person occlusion. There are two major approaches, namely bottom-up and top-down approaches. Bottom-up methods [2, 36, 40] first detect all joints of the people, and group them to each person. Top-down methods [3, 10, 37] first detect all human bounding boxes, and apply a single-person 2D pose estimation method to each person. Top-down methods generally achieve higher accuracy on traditional 2D pose benchmarks such as MSCOCO [22], but underperform on crowded scene benchmarks [21, 52] than bottom-up methods due to the human detection issues.

Recently, a few works explicitly addressed crowded scenes 2D pose estimation and reported good accuracy on crowded scene benchmarks. [21] combined top-down and bottom-up approaches using joint-candidate single person pose estimation and global maximum joints association. [4] proposed to learn scale-aware representations using high-resolution feature pyramids. [15] made a grouping process of the bottom-up approach differentiable using a graph neural network. [41] refined invisible joints' prediction using an image-guided progressive graph convolutional network.

3D human geometry estimation from crowded scenes. Several methods [29, 47, 53] have shown reasonable results on multi-person 3D benchmarks [16, 26]. However, their focus was on absolute depth estimation of each person, and few works have addressed the inter-person occlusion to estimate robust 3D geometry, such as 3D human pose (*i.e.* 3D joint coordinates) and meshes, from in-the-wild crowded scenes. XNect [27] proposed an occlusion-robust method that can be applied to crowded scenes. However, it did not focus on resolving the domain gap. It integrated 2D/3D branches into a single system and trained it on a MoCap dataset [28], which barely contains inter-person occlusions. Also, it requires a particular joint (*i.e.* neck) must be visible for human detection. On the contrary, our key idea is leveraging *external* 2D pose estimators that are not trained on MoCap data, to alleviate the domain gap between MoCap training data and in-the-wild crowd testing data. In addition, 3DCrowdNet reconstructs full 3D human pose and shape from diverse partially invisible people in crowded scenes.

ROMP [45] introduced a bottom-up method for multi-person 3D mesh recovery that can be applied to crowded scenes. It estimates a body center heatmap and a mesh parameter map, and samples each person's mesh parameters from the parameter map using center locations regressed from the heatmap. While the method provides better results on crowded scenes than previous methods, it could still suffer from the domain gap between MoCap training data and testing data from in-the-wild crowded scenes. Also, solely relying on the body center estimation to distinguish a target from others could be unstable in cases of occlusion on the body center. On the other hand, 3DCrowdNet explicitly tackles the domain gap issue with crowded-scene robust 2D poses. Also, we utilize cues from multiple 2D joint locations of the target and refine image features sampled from the locations to handle diverse inter-person occlusion, including occlusion on the body center.

2D geometry to 3D human mesh estimation. [6, 42, 43, 51] proposed methods that only take 2D geometry without images, such as 2D joint locations, for SMPL parameter regression. While the methods can benefit from 2D estimators robust to in-the-wild crowded scenes, they have two limitations. First, they cannot correct inaccurate 2D input compared to the actual person in images. Instead, they produce the most plausible outputs for the given 2D input, not the 3D pose and shape that best describes the person in images. Second, they do not benefit from image features with rich depth and 3D shape cues of a target person. The cues include subtle light reflection and shadows. 2D geometry hardly contains such cues and could lead to inaccurate 3D human mesh estimation. On the contrary, 3DCrowdNet reconstructs accurate 3D human meshes from possibly inaccurate 2D poses utilizing image features. Also, we focus on extracting the crowded scene-robust image feature of a target person using the 2D pose, rather than directly lifting 2D to 3D as the prior works.

3. 3DCrowdNet

3.1. 3DCrowdNet architecture

As shown in Figure 2, our architecture comprises a feature extractor followed by a joint-based regressor. The feature extractor is based on ResNet-50 [11], and the joint-based regressor is based on [23, 33]. Our network's output is SMPL [24] parameters, and a single person's 3D mesh is obtained by feeding the parameters to the SMPL layer.

Feature extractor. The feature extractor takes a 2D pose and an image as input. The 2D pose is 2D joint coordinates $\mathbf{P}^{\text{2D}} \in \mathbb{R}^{J\times 2}$ predicted by bottom-up off-the-shelf 2D pose estimators [2, 4]. J denotes the number of human joints, and it can vary among different 2D pose estimators. During

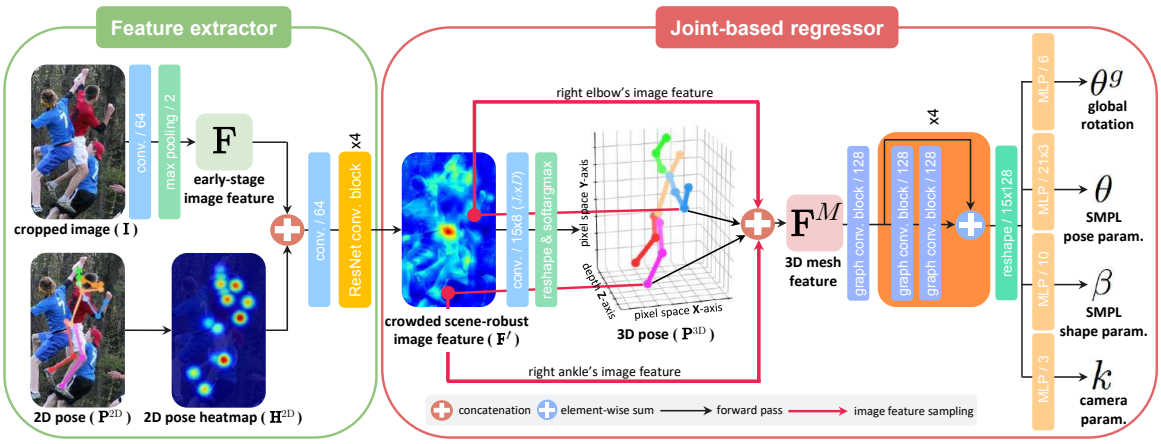


Figure 2. Overview of 3DCrowdNet. It resolves the domain gap by explicitly guiding a deep CNN to extract a crowded scene-robust feature using an off-the-shelf 2D pose estimator. Then, it distinguishes a target person from others by preserving the target's spatial activation with a joint-based regressor and regresses SMPL [24] parameters. The parameters are fed to the SMPL layer to get a 3D mesh. For simplicity, we show image feature sampling on only two joints. The numbers in network layers indicate the output channel dimension. The number in the max pooling layer indicates a stride size. The graph convolutional blocks' channel dimension is defined per joint.

training, we add realistic errors on the ground truth (GT) 2D pose following [6, 30] to mimic erroneous 2D pose outputs in test time, and the noisy 2D pose is used as our input $\mathbf{P}^{\text{2D}}$. We provide the 2D pose $\mathbf{P}^{\text{2D}}$ as a heatmap representation $\mathbf{H}^{\text{2D}} \in \mathbb{R}^{J_s \times 64 \times 64}$ to the feature extractor by making a Gaussian blob on the 2D joint coordinates. $J_s = 30$ indicates the number of joints in a superset of joint sets defined by multiple datasets. We assign don't-care values to the undefined joints and joint predictions with low confidence in inference time, by multiplying zero to the corresponding joint's heatmap. Modeling don't-care values based on the superset of joints and heatmaps enables 3DCrowdNet to perform inference from various human joint sets with a single network and handle diverse input such as 2D poses with missing joints due to truncation and occlusion.

The feature extractor uses the 2D pose heatmap $\mathbf{H}^{\text{2D}}$ of a target person as guidance and pays attention to the spatial region of a target in a crowd. First, it obtains an early-stage image feature of ResNet $\mathbf{F} \in \mathbb{R}^{C \times 64 \times 64}$ from a cropped image $\mathbf{I} \in \mathbb{R}^{3 \times 256 \times 256}$. $C = 64$ is the channel dimension, and $\mathbf{I}$ is acquired by cropping and resizing a bounding box area, derived from the 2D pose $\mathbf{P}^{\text{2D}}$. Second, it concatenates $\mathbf{F}$ and $\mathbf{H}^{\text{2D}}$ along the channel dimension. The concatenated feature is processed by a 3-by-3 convolution block, which keeps the feature's height and width but changes the channel dimension to C. Finally, the feature with C channels is fed back to the remaining part of ResNet, where the output is a crowded scene-robust image feature $\mathbf{F}' \in \mathbb{R}^{C' \times 8 \times 8}$. $C' = 2048$ is the channel dimension.

Joint-based regressor. The joint-based regressor first recovers 3D joint coordinates $\mathbf{P}^{\text{3D}} \in \mathbb{R}^{J_c \times 3}$ from $\mathbf{F}'$. $J_c = 15$ denotes the number of joints in the intersection of joint sets defined by multiple datasets. (x,y) values of $\mathbf{P}^{\text{3D}}$ are defined in a 2D pixel space, and z value of $\mathbf{P}^{\text{3D}}$ represents root joint-relative depth. A 1-by-1 convolutional layer outputs a 3D heatmap $\mathbf{H}^{\text{3D}} \in \mathbb{R}^{J_c D \times 8 \times 8}$ from $\mathbf{F}'$ after predicting a $J_c D$ dimensional 2D feature map and reshaping it to the 3D heatmap. $D = 8$ decides a descritizated size of depth. $\mathbf{P}^{\text{3D}}$ is computed from $\mathbf{H}^{\text{3D}}$, using soft-argmax operation [44]. As the soft-argmax computes continuous coordinates from a discretized grid, we observed that a heatmap with a low resolution like $\mathbf{H}^{\text{3D}}$ gives similar accuracy compared to upsampled ones, while requiring less computational costs.

Next, the joint-based regressor estimates global rotation of a person $\theta^g \in \mathbb{R}^3$, SMPL body rotation parameters $\theta \in \mathbb{R}^{21 \times 3}$, SMPL shape parameters $\beta \in \mathbb{R}^{10}$, and camera parameters $k \in \mathbb{R}^3$ for projection. First, image features per joint are sampled from $\mathbf{F}'$ using the (x,y) pixel positions of $\mathbf{P}^{\text{3D}}$. We use bilinear interpolation, since the (x,y) pixel positions are not in discretized values. The prediction confidence of $\mathbf{P}^{\text{3D}}$ is sampled from $\mathbf{H}^{\text{3D}}$ in the same manner. Second, we concatenate the sampled image features, $\mathbf{P}^{\text{3D}}$, and the prediction confidence of $\mathbf{P}^{\text{3D}}$, to attain $\mathbf{F}^M \in \mathbb{R}^{J_c \times (C'+3+1)}$. Last, we process $\mathbf{F}^M$ using a graph convolutional network (GCN), and predict θ^g, VPoser [39] latent code z, β, and k from output features of GCN with separate MLP layers. θ is decoded from z. The GCN shows faster convergence during training than an MLP network, and we think the reason lies behind the character of θ. θ is parent joint-relative joint rotations, and the GCN can exploit the human kinematic prior different from an MLP. For example, the GCN can implicitly learn the valid range of

each parent joint-relative joint leveraging the relationship between human joints.

For the graph convolutional network, we use the joint-specific graph convolution [23] that learns separate weights for each graph vertex. We define learnable weight matrices $\{W_j \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}\}_{j=1}^{J_c}$ for all joints of each graph convolution layer, where C_{in} and C_{out} denotes input and output channel dimensions, respectively. Then, the output graph feature of joint j is derived as $\mathbf{F}_j^{\text{out}} = \sigma_{\text{ReLU}}(\sum_{i \in \hat{\mathcal{N}}_j} \tilde{a}_{ji} \sigma_{\text{BN}}(W_i \mathbf{F}_i^{\text{in}}))$, where $\mathbf{F}_i^{\text{in}}$ is the input graph feature of joint i. σ_{ReLU} and σ_{BN} denotes ReLU activation function and 1D batch normalization [12], respectively. $\hat{\mathcal{N}}_j$ is defined as $\mathcal{N}_j \cup \{j\}$, where $\mathcal{N}_j$ denotes neighbors of a vertex j. $\tilde{a}_{ji}$ is an entry of the normalized adjacency matrix $\tilde{\mathbf{A}}$ at (j, i), where $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-\frac{1}{2}}$. $\mathbf{A} \in \{0, 1\}^{J_c \times J_c}$ is the adjacency matrix constructed based on the human skeleton hierarchy and fixed during the training and testing stages. The definition of the human skeleton hierarchy is depicted in the supplementary material.

3.2. Network training

The feature extractor and joint-based regressor are integrated and trained end-to-end. We use both pseudo-GT SMPL fits obtained by fitting frameworks [32, 39] and GT annotations from training datasets for supervision following [19]. Our overall objective is defined as follows:

$$L = L_{\text{pose}} + L_{\text{mesh}}, \tag{1}$$

where L_{pose} computes the L1 distance between the predicted $\mathbf{P}^{\text{3D}}$ and the (pseudo) GT, and L_{mesh} denotes the loss function for predicted SMPL parameters. L_{mesh} is defined as

$$L_{\text{mesh}} = L_{\text{param}} + L_{\text{pose}'}, \tag{2}$$

where L_{param} computes the L1 distance between the predicted θ^g, θ, and β, and the pseudo-GT parameters; $L_{\text{pose}'}$ indicates the L1 distance loss of joints regressed from predicted meshes. To supervise with 2D annotations [1, 22], predicted joints are projected by camera parameters k.

3.3. Implementation detail

PyTorch [38] is used for implementation. We initialize the weights of ResNet [11] with the pre-trained weights from [48]. It shows faster convergence during training. We use Adam optimizer [18] with a mini-batch size of 64. The initial learning rate is 10^{-4}. The model is trained for 6 epochs, and the learning rate is reduced by a factor of 10 after the 3th and 5th epochs. We use four NVIDIA RTX 2080 Ti GPUs for training, and it takes about 9 hours on average. We will release the codes for more details.

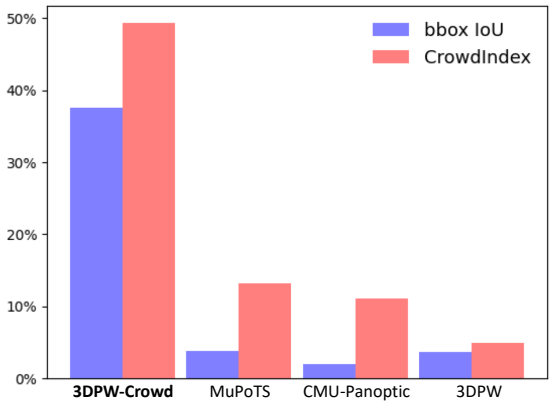


Figure 3. We curate 3DPW-Crowd, a subset of 3DPW, which has much higher bounding box IoU and CrowdIndex [21] than other 3D benchmarks. CrowdIndex measures other people's joints' ratio over each person's joints in a bounding box.

4. Experiment

4.1. Datasets

Training sets. We use Human3.6M [13], MuCo-3DHP [28], MSCOCO [22], MPII [1], and CrowdPose [21] for training. Only the training sets of the datasets are used, following the standard split protocols.

Testing sets. We report accuracy on MuPoTS [28], CMU-Panoptic [16], 3DPW [46], and 3DPW-Crowd. MuPoTS is a multi-person test benchmark captured from indoor and outdoor environments, starring 3 to 4 people. CMU-Panoptic is a large-scale multi-person dataset captured from the Panoptic studio. Following [14, 50], we pick four sequences presenting 3 to 7 people socializing each other for the evaluation. 3DPW is a widely-used 3D benchmark captured from an in-the-wild environment, and we use the test set of 3DPW following the official split protocol. 3DPW-Crowd is a subset of 3DPW and is used to evaluate the a method's robustness to in-the-wild crowded scenes. Refer to more details below about its necessity.

4.2. Evaluation protocols

Evaluation on crowded scenes: 3DPW-Crowd and CrowdPose. As CrowdPose [21] addressed, the principal obstacle of pose estimation from crowded scenes is not the number of people, but the inter-person occlusion in a crowd. Thus, MuPoTS [28] and CMU-Panoptic [16] have limitations for the evaluation on in-the-wild crowded scenes, not only because they are not in-the-wild data, but also because they show limited interaction.

To overcome the limitations, we propose 3DPW-Crowd to numerically measure a method's robustness on in-the-wild crowded scenes. It contains hugging and dancing sequences that have considerably higher average intersection over union (IoU) of bounding boxes and CrowdIndex [21]

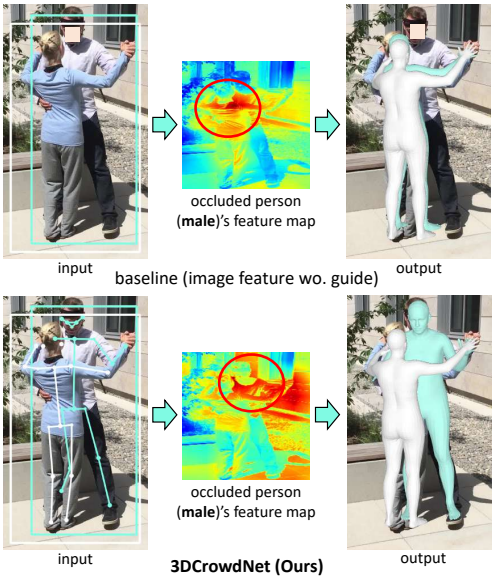


Figure 4. Comparison between the baseline that takes only image features and 3DCrowdNet. The baseline gives stronger attention to an occluding person (female) instead of an occluded person (male), and produces a wrong 3D mesh. 3DCrowdNet pays attention to the target male and recovers an accurate 3D mesh.

input feature	MPJPE↓	PA-MPJPE↓
image feature wo. guide	109.6	63.3
crowded scene-robust image feature	**85.8**	**55.8**

Table 1. Ablation on the input image features.

among 3D benchmarks as shown in Figure 3. We name the subset as 3DPW-Crowd, since it reveals the challenges of in-the-wild crowded scenes, such as overlapping bounding boxes and severe inter-person occlusion. More details about 3DPW-Crowd are in the supplementary material. We also provide extensive qualitative comparison between different methods on the test set of CrowdPose [21] in this manuscript and supplementary material.

Evaluation metrics. We report 3D pose and 3D shape evaluation metrics. For the 3D pose evaluation, we use mean per-joint position error (MPJPE), Procrustes-aligned mean per-joint position error (PA-MPJPE), and 3DPCK proposed in [26]. Following SPIN [19], we use the 3D joint coordinates regressed from a 3D mesh as predictions. For the 3D shape evaluation, we use mean per-vertex position error (MPVPE). All errors are measured after aligning root joints of GT and estimated human body meshes.

4.3. Ablation study

We carry out the ablation study on 3DPW-Crowd.We use HigherHRNet [4]'s 2D pose outputs in Table 2, 3, and 4.

Crowded scene-robust image feature. Table 1 shows the effectiveness of the crowded scene-robust image feature.The baseline network in the first row crops an image using a GT bounding box and extracts image features without any guidance as in previous methods. The significant error drop in the table proves that the 2D pose can produce crowded scene-robust image features, and the image features are critical to estimate an accurate mesh from the in-the-wild crowded scenes. We further validate our statement that the 2D pose can produce crowded scene-robust image features in Figure 4. 3DCrowdNet activates the occluded target male's spatial region, unlike the baseline network, and successfully distinguishes him from the other. As a result, 3DCrowdNet estimates an accurate mesh of the occluded target male, while the baseline predicts a mesh of the occluding female. We conclude that for the 3D mesh estimation on in-the-wild crowded scenes, the domain difference of MoCap train data is the bottleneck, and our idea to exploit the robustness of 2D pose estimators, which do not use MoCap train data, is valid.

parameter regressor type	MPJPE↓	PA-MPJPE↓
SPIN-style regressor	89.0	59.5
joint-based regressor (Ours)	**85.8**	**55.8**

Table 2. Ablation of the regressor types.

sampling area	MPJPE↓	PA-MPJPE↓
whole feature map	89.1	57.8
5-by-5 grid around point	88.2	57.6
point (Ours)	**85.8**	**55.8**

Table 3. Ablation on the sampling area of image features.

Joint-based regressor. Table 2 shows that the joint-based regressor outperforms the SPIN [19]-style regressor, the dominant model-based approach in the current literature, on 3DPW-Crowd. The results prove that preserving the spatial activation of a target person in a deep CNN feature map is essential. SPIN-style regressor shows lower accuracy, since it makes a target person's feature indistinguishable from others by collapsing the spatial information with a global average pooling. We further validate our argument in Table 3. Originally, our joint-based regressor samples deep image features from (x,y) positions of the predicted 3D pose. When we enlarge the sampling area, the errors increase. Especially, when the joint-based regressor uses features sampled from the whole feature map, which are the same feature of the SPIN-style regressor, MPJPE becomes similar to that of the SPIN-style regressor. It indicates that most of the accuracy gain in Table 2 is not from a better network architecture, such as GCN, but from the preservation of the target person's spatial activation. Keeping an appropriate sampling area to less involve non-target people's image features is important to estimate a robust human mesh from in-the-wild crowded scenes.

We also verify the effectiveness of estimating a 3D pose

estimation target	MPJPE↓	PA-MPJPE↓
2D pose	88.3	56.4
3D pose (Ours)	**85.8**	**55.8**

Table 4. Ablation on the intermediate estimation target of the joint-based regressor during training and testing.

method	MPJPE↓	PA-MPJPE↓	MPVPE↓
SPIN [19]	121.2	69.9	144.1
Pose2Mesh [6]	124.8	79.8	149.5
I2L-MeshNet [31]	115.7	73.5	162.0
ROMP [45]∗	104.8	63.9	127.8
3DCrowdNet (Ours)	86.8	56.1	109.7
3DCrowdNet (Ours)∗	**85.8**	**55.8**	**108.5**

Table 5. Comparison on 3DPW-Crowd between 3DCrowdNet and previous methods. We evaluate other methods with their codes and pre-trained models. ∗ means using CrowdPose [21] for training.

instead of a 2D pose in Table 4. The clear accuracy improvement proves that the depth information can be reliably estimated from a 2D pose and image features, and it is beneficial for the accuracy of the final mesh estimation.

4.4. Comparison with state-of-the-art methods

Unless indicated, our 3DCrowdNet is not trained on a CrowdPose [21] train set in Table 5, 6, and 7. Also, we use less or similar training data than other methods, and the details are in the supplementary material.

3DPW-Crowd. We compare our 3DCrowdNet with [6, 19, 31, 45] in Table 5. They are recent state-of-the-art 3D human mesh estimation methods on 3DPW, and publicly released the codes for evaluation. We make several observations. First, our approach outperforms SPIN [19], which takes only the image feature as input and performs a global average pooling on the deep CNN feature map. The result is coherent with the results in Table 1 and 2 of our ablation studies. Next, 3DCrowdNet outperforms ROMP [45], a bottom-up method for multi-person 3D mesh estimation. While ROMP achieves higher accuracy than other methods, we think it still suffers from the domain gap issue. For example, it needs to learn how to distinguish body centers of people under diverse inter-person occlusion, but MoCap datasets they used rarely contain such data. On the other hand, 3DCrowdNet explicitly resolves the domain gap using 2D pose input and produces accurate 3D meshes.

Last, 3DCrowdNet defeats Pose2Mesh [6], a method that can also benefit from crowded-scene robust 2D poses. We used the same 2D pose predictions of [4] for Pose2Mesh and 3DCrowdNet. The result validates 3DCrowdNet's two strengths over Pose2Mesh. First, 3DCrowdNet recovers a 3D mesh that best describes a target person, using rich depth and shape cues in images. On the contrary, Pose2Mesh produces the most plausible 3D mesh for a given 2D pose, and the accuracy depends on it. Figure 6 shows that 3DCrowdNet recovers accurate 3D meshes, even when a 2D pose is inaccurate. Second, 3DCrowdNet can handle missing joints of 2D pose predictions due to occlusion and truncation owing to don't-care modeling based on the 2D pose's heatmap introduced in Section 3.1. Pose2Mesh takes the 2D pose as coordinates and cannot cope with the missing joints, common in in-the-wild crowded scenes. Please also refer to the qualitative comparison in the supplementary material.

method	3DPCK↑	
	All	Matched
SMPLify-X [39] / OpenPose [2]	62.8	68.0
HMR [17] / OpenPose [2]	66.0	70.9
HMR [17] / Mask R-CNN [10]	65.6	68.6
Jiang *et al*. [14]	69.1	72.2
3DCrowdNet (Ours) / OpenPose [2]	70.2	70.9
3DCrowdNet (Ours) / HigherHRNet [4]	**72.7**	**73.3**

Table 6. Comparison on MuPoTS [28] between 3DCrowdNet and previous methods. The numbers denote 3DPCK for all annotations (All) and annotations matched to a prediction (Matched), and are brought from [14]. The method names beside [2,4,10] indicate the source of bounding boxes and 2D pose input.

method	Haggl.	Mafia	Ultim.	Pizza	Mean
Zanfir *et al*. [49]	140.0	165.9	150.7	156.0	153.4
Zanfir *et al*. [50]	141.4	152.3	145.0	162.5	150.3
Jiang *et al*. [14]	129.6	133.5	153.0	156.7	143.2
ROMP [45]	111.8	**129.0**	148.5	149.1	134.6
3DCrowdNet (Ours)	**109.60**	135.9	**129.8**	**135.6**	**127.6**

Table 7. Comparison on CMU-Panoptic [16]. The numbers denote MPJPE. We follow the evaluation protocol of Jiang *et al*. [14].

MuPoTS. Table 6 compares our 3DCrowdNet with methods that recover a 3D mesh. It outperforms all the previous methods. Note that the second and fifth rows prove that 3DCrowdNet's high accuracy on the crowded scenes is not simply attributed to better localization derived from bottom-up 2D poses. While 3DCrowdNet and HMR use the same 2D poses of OpenPose [2], HMR utilizes the 2D pose only to get a bounding box, and 3DCrowdNet additionally uses the 2D pose to guide a feature extractor to extract crowded scene-robust image features. Leveraging more information in given input is natural, and leads to better accuracy.

CMU-Panoptic. Table 7 shows that our 3DCrowdNet significantly outperforms previous 3D human pose and shape estimation methods on CMU-Panoptic. The result demonstrates that the proposed 3DCrowdNet can perform competitively on crowded scenes with daily social activities. Note that no data from CMU-Panoptic are used for training.

3DPW. Table 8 shows that 3DCrowdNet achieves state-of-the art accuracy in general in-the-wild scenes. The result validates that 3DCrowdNet is robust to diverse challenges of in-the-wild scenes, although our method is designed to target crowded scenes. The second row of Figure 5 supports

Figure 5. Qualitative comparison on a CrowdPose [21] test set with SPIN [19] and ROMP [45]. We highlighted their representative failure cases with red circles. The order of 3D meshes is manually assigned.

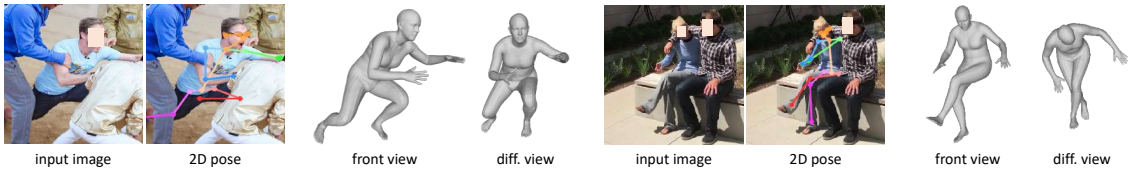


Figure 6. Visualization of a 3D mesh from different viewpoints. 3DCrowdNet effectively resolves the depth ambiguity using the cues of 2D pose input and image features.

method	MPJPE↓	PA-MPJPE↓	MPVPE↓
HMR [17]	130	76.7	-
GraphCMR [20]	-	70.2	-
SPIN [19]	96.9	59.2	116.4
I2L-MeshNet [31]	93.2	57.7	110.1
Pose2Mesh [6]	89.5	56.3	105.3
Song *et al*. [43]	-	55.9	-
Fang *et al*. [7]	85.1	54.8	-
TUCH [34]	84.9	55.5	-
ROMP [45]	91.3	54.9	108.3
3DCrowdNet (Ours)	**81.7**	**51.5**	**98.3**

Table 8. Comparison on 3DPW [46] between 3DCrowdNet and state-of-the-art methods of 3D human mesh estimation from a single image. We compare methods that do not use 3DPW train set during training for the fair comparison.

our statement, which shows 3DCrowdNet's robustness to truncation and occlusion in in-the-wild images.

We provide the qualitative comparison with SPIN [19] and ROMP [45] in Figure 5. Apparently, 3DCrowdNet produces much more robust 3D meshes on in-the-wild crowded scenes. SPIN predicts a swapped leg pose (top), fails to distinguish different people in the overlapping bounding boxes (middle), and misses the right leg's pose due to inter-person occlusion (bottom). ROMP produces an inaccurate pose for a person under occlusion with similar appearances (top), misses a target whose body center (*i.e.* torso) is invisible (middle), and estimate an inaccurate global rotation of a target due to occlusion by a nearby person with similar appearance (bottom). Please also refer to more extensive qualitative comparison with [6, 19, 31, 45] and failure cases of 3DCrowdNet in the supplementary material.

5. Conclusion

We present 3DCrowdNet, the first single image-based 3D human mesh estimation system that explicitly targets in-the-wild crowded scenes. It extracts crowded-scene robust image features of a target person, and effectively distinguishes the target from others. We guide a deep CNN to pay attention to the target using a 2D pose, which is robust to the domain gap between MoCap training data and crowd testing data. The joint-based regressor preserves the spatial activation of the target, and effectively excludes non-target people's image features. We show that 3DCrowdNet highly outperforms previous methods on in-the-wild crowded scenes both quantitatively and qualitatively. 3DCrowdNet could be a baseline for future image-based methods that target crowded scenes owing to the simple yet effective implementation.

Acknowledgements. This work was supported in part by IITP grant funded by the Korea government (MSIT) [No. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 3, 7

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 3

[4] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 3, 6, 7

[5] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *CVPR*, 2021. 2

[6] 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 *ECCV*, 2020. 2, 3, 4, 7, 8

[7] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3D human pose by watching humans in the mirror. In *CVPR*, 2021. 8

[8] Thomas Gatt, Dylan Seychell, and Alexiei Dingli. Detecting human abnormal behaviour through a video generated model. In *IEEE ISPA*, 2019. 1

[9] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 2

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 3, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[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. *IEEE TPAMI*, 2014. 2, 5

[14] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 5, 7

[15] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *ECCV*, 2020. 3

[16] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic Studio: A massively multiview system for social interaction capture. *IEEE TPAMI*, 2017. 2, 3, 5, 7

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 7, 8

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[19] 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 *ICCV*, 2019. 1, 2, 5, 6, 7, 8

[20] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 8

[21] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 2, 3, 5, 6, 7, 8

[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 *ECCV*, 2014. 2, 3, 5

[23] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3D human pose estimation. In *ECCV*, 2020. 3, 5

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 2, 3, 4

[25] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 2

[26] 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 *3DV*, 2017. 2, 3, 6

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

[28] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *3DV*, 2018. 2, 3, 5, 7

[29] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image. In *ICCV*, 2019. 3

[30] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. PoseFix: Model-agnostic general human pose refinement network. In *CVPR*, 2019. 4

[31] 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 *ECCV*, 2020. 7, 8

[32] Gyeongsik Moon and Kyoung Mu Lee. NeuralAnnot: Neural annotator for in-the-wild expressive 3D human pose and

mesh training sets. *arXiv preprint arXiv:2011.11232*, 2020. 5

[33] Gyeongsik Moon and Kyoung Mu Lee. Pose2pose: 3D positional pose-guided 3D rotational pose prediction for expressive 3d human pose and mesh estimation. *arXiv preprint arXiv:2011.11534*, 2020. 3

[34] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *CVPR*, 2021. 8

[35] Neeti Narayan, Nishant Sankaran, Srirangaraj Setlur, and Venu Govindaraju. Re-identification for online person tracking by modeling space-time continuum. In *CVPR workshop*, 2018. 1

[36] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative Embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017. 3

[37] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 3

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS*, 2017. 5

[39] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 4, 5, 7

[40] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 3

[41] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *ECCV*, 2020. 2, 3

[42] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *BMVC*, 2020. 3

[43] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 3, 8

[44] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017. 4

[45] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *ICCV*, 2021. 3, 7, 8

[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 *ECCV*, 2018. 2, 5, 8

[47] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. HMOR: Hierarchical multi-person ordinal relations for monocular multi-person 3D pose estimation. In *ECCV*, 2020. 3

[48] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 5

[49] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 7

[50] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. In *NeurIPS*, 2018. 5, 7

[51] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3D human shape and pose from dense body parts. *IEEE TPAMI*, 2020. 3

[52] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2Seg: Detection free human instance segmentation. In *CVPR*, 2019. 3

[53] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: Single-shot multi-person absolute 3D pose estimation. In *ECCV*, 2020. 3