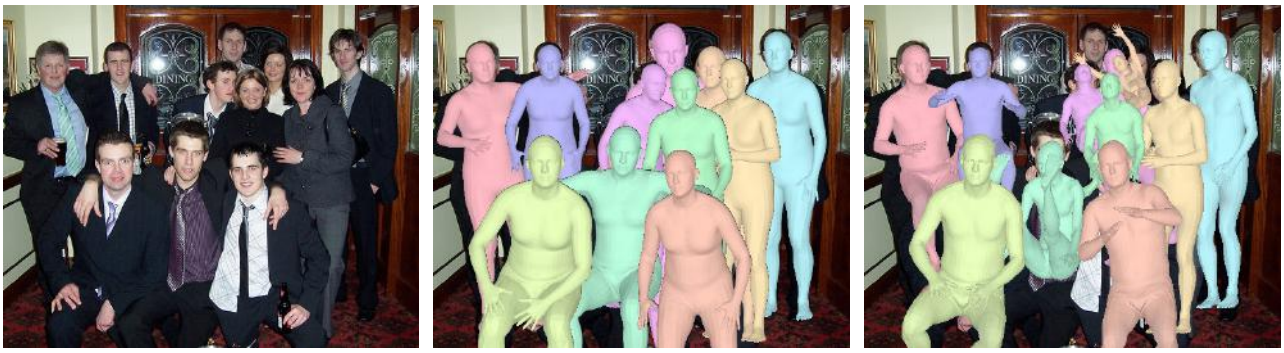
3DCrowdNet: 2D Human Pose-Guided 3D Crowd Human Pose and Shape Estimation in the Wild

Hongsuk Choi Gyeongsik Moon JoonKyu Park Kyoung Mu Lee

ECE & ASRI, Seoul National University, Korea
{redarknight,mks0601,jkpark0825,kyoungmu}@snu.ac.kr



input image **3DCrowdNet (Ours)** SPIN

Figure 1: Existing single-person 3D human pose and shape estimation methods, such as SPIN [21], give inaccurate results on in-the-wild crowd scenes. Our 3DCrowdNet produces much more robust outputs for each individual, effectively leveraging the image feature and the 2D human pose outputs from off-the-shelf 2D pose estimators.

Abstract

Recovering accurate 3D human pose and shape from in-the-wild crowd scenes is highly challenging and barely studied, despite their common presence. In this regard, we present 3DCrowdNet, a 2D human pose-guided 3D crowd pose and shape estimation system for in-the-wild scenes. 2D human pose estimation methods provide relatively robust outputs on crowd scenes than 3D human pose estimation methods, as they can exploit in-the-wild multi-person 2D datasets that include crowd scenes. On the other hand, the 3D methods leverage 3D datasets, of which images mostly contain a single actor without a crowd. The train data difference impedes the 3D methods' ability to focus on a target person in in-the-wild crowd scenes. Thus, we design our system to leverage the robust 2D pose outputs from off-the-shelf 2D pose estimators, which guide a network to focus on a target person and provide essential human articulation information. We show that our 3DCrowdNet outperforms previous methods on in-the-wild crowd scenes. We will release the codes.

1. Introduction

Various methods have been proposed to recover 3D human pose and shape from a single image. However, the 3D pose and shape estimation from in-the-wild crowd scenes has been barely studied, despite their common presence.

Most of the previous 3D pose and shape estimation methods [9, 19, 21, 39] first crop an image using a bounding box of a target person detected from off-the-shelf human detectors [11, 42]. Then they take the target person's cropped image and regress the human model parameters, such as SMPL [26]. While the previous methods achieve high accuracy when provided with accurate human bounding boxes and little inter-person occlusion, they tend to output poor results on in-the-wild crowd scenes as depicted in Figure 1. In-the-wild crowd scenes involve overlapping and inaccurate human bounding boxes, which often include other people. The noisy bounding boxes confuse the network on which to focus. Inter-person occlusion in the crowd scenes hinders accurate 3D pose prediction and causes swapped joint prediction between people.

On the other hand, 2D human pose estimation meth-

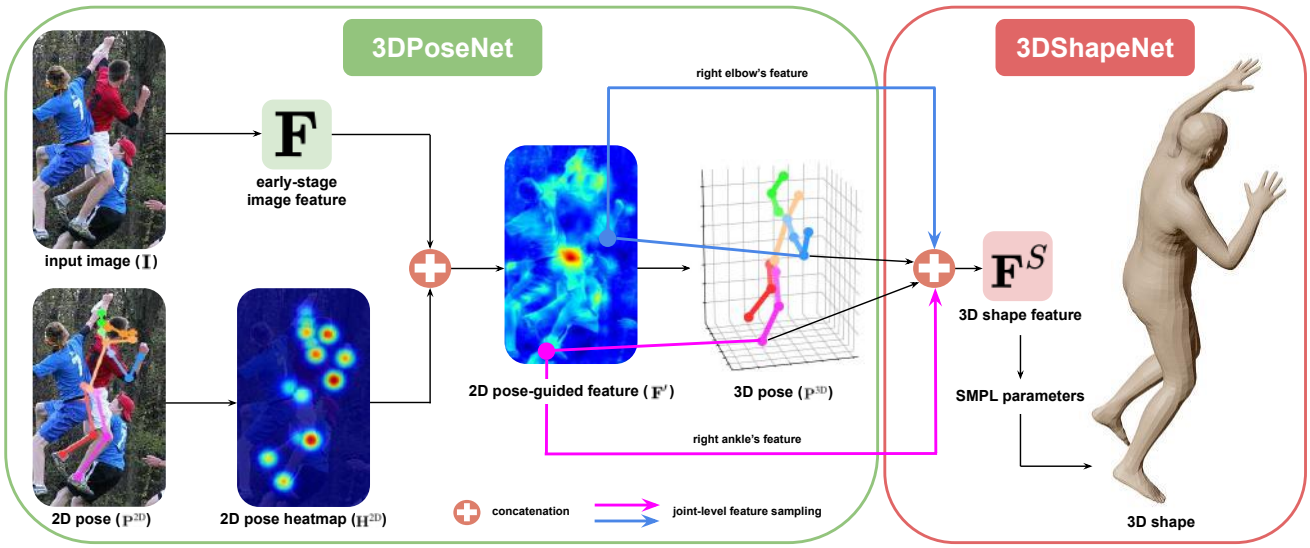


Figure 2: The overall architecture of 3DCrowdNet, which consists of 3DPoseNet and 3DShapeNet. 3DPoseNet produces 2D pose-guided feature that has high activation on a target person. Then it estimates 3D joint coordinates from it. 3DShapeNet samples joint-level feature from the 2D pose-guided feature based on (x,y) positions of the predicted 3D joint coordinates. Then it outputs SMPL parameters, which are decoded to a 3D shape.

ods tend to suffer less from the inaccurate human bounding boxes and inter-person occlusion. We argue that the difference is attributed to the composition of train data. 3D human pose and shape estimation methods typically leverage 3D human pose train datasets [15, 27] to learn depth information from images. However, to attain accurate ground truth (GT) 3D annotations via multi-view geometry algorithm, many 3D datasets are captured from a multi-view studio starring a single person without a crowd. The perfect human bounding boxes and no inter-person occlusion are guaranteed for accurate 3D annotations, but the images are far from in-the-wild crowd scenes. On the contrary, 2D human pose estimation methods are trained on diverse in-the-wild data [1, 24, 51]. The 2D datasets contain images with heavily overlapping bounding boxes and inter-person occlusion, where human manually annotates the data. While the 3D methods can use the in-the-wild 2D data in a weakly-supervised manner, their strong reliance on 3D data leads to low pose accuracy on crowd scenes.

In this regard, we propose 3DCrowdNet, a 2D human pose-guided system for 3D human pose and shape estimation from in-the-wild crowd scenes. This study is the first to explicitly tackle the 3D crowd human pose and shape estimation from a single image to the best of our knowledge. 3DCrowdNet consists of 3DPoseNet and 3DShapeNet, and the inference process proceeds in three steps. First, we leverage off-the-shelf bottom-up 2D pose estimators, which first detect all human joints and group them to each person, to obtain crowd-scene robust 2D pose outputs and bounding boxes of each person. The bounding boxes are computed by enlarging the 2D pose boundaries and used to acquire each person's image feature. The bottom-up 2D pose estimators allow us to obtain the 2D poses and bounding boxes simultaneously. Second, 3DPoseNet concatenates the 2D pose output and the image feature from a cropped person to guide a network to focus on the person who is the target. The 2D pose output activates the target person's image feature area and encourages the network to disentangle the target person from other people included in a bounding box. In addition, it provides essential human articulation information to the 3DCrowdNet. Last, 3DPoseNet refines the possibly noisy 2D pose output, and 3DShapeNet estimates SMPL parameters. 3DShapeNet is a joint-based regressor, which extracts joint-level feature from the 2D pose-guided image feature based on the output from 3DPoseNet. Figure 2 describes the overall architecture.

The experimental results show that the proposed 3DCrowdNet significantly outperforms the previous 3D human pose and shape estimation methods on in-the-wild crowd scenes. Also, it achieves state-of-the-art accuracy on multi-person 3D benchmarks [18, 29] and in-the-wild benchmark [49]. Extensive ablation experiments are carried out to validate the system, and would give insight for future works of 3D human pose and shape estimation from in-the-wild crowd scenes.

Our contributions can be summarized as follows.

- We present 3DCrowdNet, which firstly targets in-the-wild crowd scenes for 3D human pose and shape estimation from a single image. The proposed system leverages the crowd-scene robust 2D pose outputs to guide a network to focus on a target person.

- We provide extensive ablation experiments to justify the design choices of our system. 3DCrowdNet outperforms previous methods on both in-the-wild crowd scenes and conventional 3D benchmarks.

2. Related works

Crowd scenes 2D human pose estimation. Early works of multi-person 2D human pose estimation did not explicitly target crowd scenes. However, their methods are related to diverse challenges from crowd scenes, such as overlapping human bounding boxes, human detection error, and inter-person occlusion. Typically, there are two approaches, namely bottom-up and top-down approaches. Bottom-up methods [2,13,34,40] first detect all joints of the people, and group them to each person. Top-down methods [5,8,11,36] first detect all human bounding boxes, and apply a single-person 2D pose estimation method to each person. Top-down methods generally achieve lower error on traditional 2D pose benchmarks such as MSCOCO [24], but underperform on crowd scene benchmarks [23,41,57] than bottom-up methods due to overlapping human bounding boxes and human detection error.

Recently, many works explicitly address crowd scenes 2D human pose estimation or report accuracy on crowd scene benchmarks. Li *et al*. [23] combined top-down and bottom-up approaches using joint-candidate single person pose estimation and global maximum joints association. Cheng *et al*. [6] proposed to learn scale-aware representations using high-resolution feature pyramids. It is the current state-of-the-art method among bottom-up methods on MSCOCO [24], and achieves the best accuracy on CrowdPose [23], the popular 2D crowd scene benchmark. Jin *et al*. [17] made a grouping process of the bottom-up approach differentiable using a graph neural network. Qui *et al*. [41] suggested refining invisible joints' prediction using an image-guided progressive graph convolutional network.

Multi-person 3D human pose estimation. Few multi-person 3D human pose estimation works have focused on crowd scenes 3D human pose estimation. Rogez *et al*. [43,44] introduced a top-down method called LCR-Net, which consists of localization, classification, and regression components. Mehta *et al*. [29] introduced a bottom-up approach using an occlusion-robust pose-map (ORPM) formulation. Moon *et al*. [30] and Wang *et al*. [50] proposed fully learning-based multi-person 3D human pose estimation methods, which focus on recovering absolute depth of each person. Zhen *et al*. [58] exploited the method of Cao *et al*. [2], and incorporated absolute depth prediction for grouping joints to each person. Recently, Chen *et al*. [4] explicitly targeted crowd scenes based on multi-view geometry. They proposed a graph model for fast cross-view matching and MAP optimization for robust 3D crowd reconstruction. XNect *et al*. [28] extended Mehta *et al*. [29] by proposing a novel backbone SelecSLS Net and a 3D pose encoding map that improved ORPM.

Although XNect [28] is an occlusion robust method, it requires multiple pre-training steps on 2D and 3D datasets [1, 24, 29], and a certain joint (*i.e.* neck) must be visible for human detection. On the contrary, our system can be end-to-end trained without complicated pre-training steps, and reconstructs full 3D human pose and shape from diverse partially invisible people in crowd scenes.

3D human pose and shape estimation in crowd scenes. Most of the previous 3D human pose and shape estimation methods [9,10,16,19,21,22,32,35,39,48,53] take images as input and regress a 3D human shape. To estimate accurate depth information, they necessarily rely on 3D human pose datasets [15,27], which are captured from a studio environment far from in-the-wild crowd scenes. Also, they are generally based on a top-down approach, which inevitably suffers from overlapping bounding boxes and human detection error. As a result, their prediction often misses an occluded person or reveals swapped joints between people, due to overlapping bounding boxes and inter-person occlusion in in-the-wild crowd scenes.

To address such issues, a pioneering paper by Zanfir *et al*. [55] proposed a bottom-up system for multi-person 3D human pose and shape estimation from a single image. It simultaneously estimates 2D and 3D poses, and resolves the joint grouping problem via a binary integer programming based on 2D pose prediction scores. Then, SMPL parameters are regressed from the predicted 3D pose based on an auto-encoding scheme. Different from the method of Zanfir *et al*. [55], 3DCrowdNet can easily leverage off-the-shelf state-of-the-art 2D pose estimators. In addition, our system infers a 3D shape from not only a 3D pose but also image feature, which are crucial for recovering accurate shape.

Guler *et al*. [10] presented HoloPose that utilizes joint-level image feature based on its' 2D joint predictions. In the case of a 2D joint prediction with low confidence, the method estimates the corresponding joint angle as an average of neighbor joint angles or fills it with the resting pose if neighbor joints also have low confidence. However, crowd scenes frequently contain heavy occlusion that brings many 2D joint predictions with low confidence, and these heuristics are likely to malfunction. On the contrary, our 3DShapeNet samples every joint-level feature from image feature and can benefit from a data-driven learning strategy.

3D human pose and shape estimation from 2D geometry. Recently, [7,45,46,56] proposed systems that only take 2D geometry information, such as 2D joint locations, for SMPL parameter regression. While the systems can benefit from in-the-wild robust 2D estimators, discarding image feature is abandoning rich 3D depth and shape cues in images. In addition, relying solely on 2D geometry drives a network to produce the most plausible outputs for the given input,

not the 3D pose and shape that best describes the person in images. On the contrary, our 3DCrowdNet leverages appearance and shadow information related to shape and depth in images, and reconstructs 3D pose and shape that corresponds to the target person.

3. 3DCrowdNet

3.1. Preparing 2D human pose input

Our proposed system uses 2D joint coordinates $\mathbf{P}^{\text{2D}} \in \mathbb{R}^{J\times 2}$ predicted by bottom-up off-the-shelf 2D human pose estimators [2, 6]. J denotes the number of human joints, and it varies among different 2D human pose estimators. The capability to exploit off-the-shelf 2D human pose estimators is advantageous, since our system can benefit from the improving 2D methods without re-training our system. During training, we mimic the 2D human pose outputs $\mathbf{P}^{\text{2D}}$ by adding realistic errors on the GT (Ground Truth) 2D pose following [3, 7, 31].

We generate a 2D pose heatmap $\mathbf{H}^{\text{2D}} \in \mathbb{R}^{J_s\times H\times W}$ from $\mathbf{P}^{\text{2D}}$ by making a Gaussian blob on the 2D coordinates. H and W denote the height and width of the image feature to be concatenated. J_s indicates the number of joints in a superset of joints sets defined by multiple datasets. We assign don't-care values to the undefined joints and joint predictions with low confidence in inference time, multiplying zero to the corresponding joint heatmap.

3.2. 3DPoseNet

3DPoseNet estimates 3D joint coordinates from the 2D joint heatmaps and an image.

2D human pose-guided image feature. 3DPoseNet uses a modified version of ResNet [12] to extract 2D human pose-guided image feature. First, it obtains early-stage image feature $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ from a cropped image $\mathbf{I} \in \mathbb{R}^{3\times 4H\times 4W}$ using the first convolution and max-pooling layers of ResNet. C is the feature channel dimension, and H and W are a quarter of the height and width of $\mathbf{I}$. The cropped image $\mathbf{I}$ is acquired using a bounding box derived from the 2D pose $\mathbf{P}^{\text{2D}}$. Then, it concatenates $\mathbf{F}$ and the 2D pose heatmap $\mathbf{H}^{\text{2D}}$ 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 2D human pose-guided image feature $\mathbf{F}' \in \mathbb{R}^{C'\times H/4\times W/4}$. Different from the original ResNet output feature, $\mathbf{F}'$ has more activation on the corresponding 2D joint locations of $\mathbf{P}^{\text{2D}}$. The 2D human pose-guided image feature $\mathbf{F}'$ is essential for crowd scenes 3D pose and shape estimation, since it disentangles the target person's image feature from others.

3D joint coordinates. 3DPoseNet recovers 3D joint coordinates $\mathbf{P}^{\text{3D}} \in \mathbb{R}^{J_c\times 3}$ from the 2D human pose-guided feature $\mathbf{F}'$. J_c denotes the number of joints that are most common among multiple datasets. (x,y) values of $\mathbf{P}^{\text{3D}}$ are defined in a 2D image coordinate space, and z value of $\mathbf{P}^{\text{3D}}$ represents root joint-relative depth. 3DPoseNet processes $\mathbf{F}'$ by a 1-by-1 convolution block, and outputs a 3D heatmap $\mathbf{H}^{\text{3D}} \in \mathbb{R}^{J_c\times H/4\times W/4}$ after reshaping the tensor. $\mathbf{P}^{\text{3D}}$ is computed from $\mathbf{H}^{\text{3D}}$ using the soft-argmax operation [47].

3.3. 3DShapeNet

3DShapeNet estimates global rotation of a person $\theta^g \in \mathbb{R}^3$, SMPL parameters $\theta \in \mathbb{R}^{21\times 3}$, $\beta \in \mathbb{R}^{10}$, and camera parameters $k \in \mathbb{R}^3$ for projection. θ and β are pose and shape parameters, respectively. β and k are regressed from $\mathbf{F}'$ after spatially averaging it, followed by a separate fully connected layer.

Joint-based regressor. θ^g and θ are estimated from $\mathbf{F}'$, $\mathbf{P}^{\text{3D}}$, and 3D joint prediction confidence of $\mathbf{P}^{\text{3D}}$ from $\mathbf{H}^{\text{3D}}$. We design 3DShapeNet as a joint-based model that preserves the 2D pose cues for the target person in crowd scenes. The joint-based model [10, 56] utilizes joint-level image feature to estimate joint angles such as SMPL pose parameters. First, 3DShapeNet samples image feature per joint from $\mathbf{F}'$ using the (x,y) position of $\mathbf{P}^{\text{3D}}$. Second, it concatenates the sampled joint-level image feature, $\mathbf{P}^{\text{3D}}$, and the 3D joint prediction confidence, to attain $\mathbf{F}^S \in \mathbb{R}^{J_c\times(C'+3+1)}$. Last, 3DShapeNet processes $\mathbf{F}^S$ using a graph convolution network, and predicts θ^g and θ by a separate fully connected layer.

Graph convolution. For the graph convolution network, we use joint-specific graph convolution 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 [14], 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. $\mathbf{D}$ is a diagonal matrix of $\mathbf{A}+\mathbf{I}$. 3DShapeNet adopts the network architecture of Liu *et al.* [25], which consists of one graph convolution block and four graph residual blocks. We empirically observed that the joint-specific graph convolution provides faster convergence on test accuracy than fully connected layers.

3.4. Network Training

3DPoseNet and 3DShapeNet are integrated and trained in an end-to-end manner. We use pseudo-GT SMPL fits obtained from a fitting framework, in addition to GT annotations from train datasets following [7, 21, 32, 53]. The loss function for 3DPoseNet is defined as follows:

$$L_{\text{pose}} = L_{\text{coord}}^{\text{3DPoseNet}}, \tag{1}$$

where $L_{\text{coord}}^{\text{3DPoseNet}}$ denotes the L1 distance between the predicted $\mathbf{P}^{\text{3D}}$ and the GT. Owing to the soft-argmax operation [47], $L_{\text{coord}}^{\text{3DPoseNet}}$ allows the 2D and 3D mixed data supervision.

To supervise 3DShapeNet, following [7, 19, 21, 32] we regress 3D pose $\hat{\mathbf{P}}^{\text{3D}} \in \mathbb{R}^{J_s \times 3}$ from the 3D shape decoded from θ^g, θ, and β. Camera parameters k is used to project $\hat{\mathbf{P}}^{\text{3D}}$ to image space, to produce $\hat{\mathbf{P}}^{\text{2D}} \in \mathbb{R}^{J_s \times 2}$. The loss function for 3DShapeNet is defined as follows:

$$L_{\text{shape}} = L_{\text{param}} + L_{\text{coord}}^{\text{3DShapeNet}}, \tag{2}$$

where L_{param} denotes the L1 distance between the predicted θ^g, θ, and β, and the pseudo-GT parameters; $L_{\text{coord}}^{\text{3DShapeNet}}$ includes the L1 distance loss of $\hat{\mathbf{P}}^{\text{3D}}$ and $\hat{\mathbf{P}}^{\text{2D}}$.

4. Implementation detail

PyTorch [37] is used for implementation. We initialize the weights of ResNet [12] with the pre-trained weights from Xiao *et al*. [52]. It proved to be effective on faster convergence during training. The weights of the whole model are updated by the Adam optimizer [20] with a mini-batch size of 64. The initial learning rate is 10^{-4}. The model is trained for six epochs, and the learning rate is reduced by a factor of 10 after the 3th and 5th epochs. The cropped image is resized to 256×256. 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 implementation details.

5. Experiment

5.1. Datasets

Training sets. We use Human3.6M [15], MuCo-3DHP [29], MSCOCO [24], MPII [1], and CrowdPose [23] for training. Only the training sets of the datasets are used, following the standard split protocols. To generate pseudo-GT of SMPL parameters, we use NeuralAnnot [33].

Testing sets. We report accuracy performance on MuPoTS [29], CMU-Panoptic [18], and 3DPW [49]. 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 [16, 55], 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 protocols. More details are in the supplementary material.

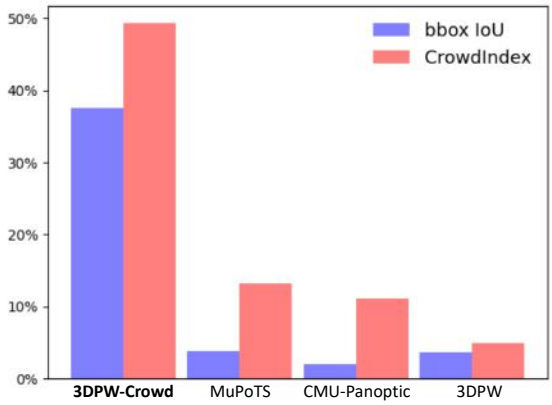


Figure 3: 3DPW-Crowd, a subset of 3DPW [49] validation set, has much higher bounding box IoU and CrowdIndex [23] than other datasets. CrowdIndex measures other people's joints' ratio over each person's joints in a bounding box. The statistics support that 3DPW-Crowd is a suitable benchmark for the evaluation on in-the-wild crowd scenes.

5.2. Evaluation protocols

Evaluation on crowd scenes. As Zhang *et al*. [23] addressed, the principal obstacle of crowd scenes pose estimation is not the number of people, but the inter-person occlusion due to diverse interaction between people. Thus, MuPoTS [29] and CMU-Panoptic [18] have limitations for the evaluation on in-the-wild crowd scenes, not only because they are not in-the-wild data, but also because they show limited interaction between people.

In this regard, we provide a numerical evaluation on a subset of 3DPW [49] validation set that shows substantial interaction, to investigate the method's robustness on in-the-wild crowd scenes thoroughly. The subset contains hugging and dancing sequences that have considerably higher average intersection over union (IoU) of bounding boxes and CrowdIndex [23] than other datasets as shown in Figure 3. We name the subset as 3DPW-Crowd, since it reveals the challenges of in-the-wild crowd scenes, such as overlapping bounding boxes and severe inter-person occlusion. More details about 3DPW-Crowd are in the supplementary material. We also show extensive qualitative results on the test set of CrowdPose [23].

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 Mehta *et al*. [27]. Following [7, 19, 21, 32], we use the 3D joint coordinates regressed from the final 3D shape as

Table 1: Comparison between networks taking different input feature on 3DPW-Crowd. For our 2D pose-guided system, we provide results using 2D pose outputs from [2, 6]. The results show that the 2D pose-guided feature from the bottom-up 2D pose estimators [2, 6] is crucial for robust estimation on crowd scenes.

input feature	MPJPE	PA-MPJPE
image feature wo. guide	109.6	63.3
2D pose [2]-guided feature	87.6	58.5
2D pose [6]-guided feature	**85.6**	**55.4**

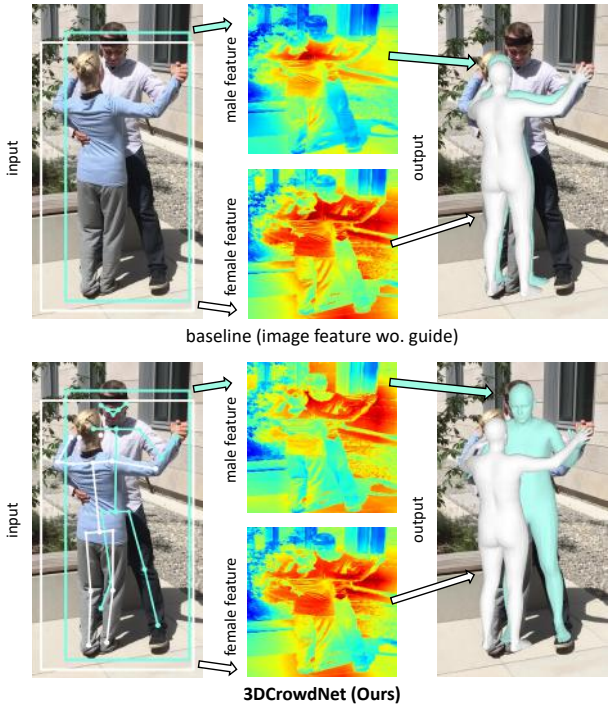


Figure 4: Qualitative comparison between 3DCrowdNet and the baseline that takes only image feature as input. A 2D pose activates the target person's area in the image feature in our system and encourages the network to disentangle the target person from other people in a bounding box.

predictions. For the 3D shape evaluation, we use mean per vertex position error (MPVPE).

5.3. Ablation study

We carry out the ablation study on 3DPW-Crowd to validate our method's robustness on crowd scenes.

Effect of 2D human pose-guided feature. Table 1 shows that the 2D human pose-guided image feature's effectiveness is definite on crowd scenes. The first-row network, the baseline, crops an image using a GT bounding box and extracts image feature without a 2D pose guide. MPJPE and PA-MPJPE drop by 24.0mm and 7.9mm respectively, when 3DCrowdNet guides the image feature with 2D pose outputs from Cheng *et al*. [6], the state-of-the-art bottom-up 2D pose estimator. The result proves that the crowd-scene robust 2D pose outputs guide a network to focus on a target person. Figure 4 illustrates how the image feature is affected by the 2D pose in our system. Unlike the baseline network, our 3DCrowdNet activates the target person's region and successfully disentangles different people who have heavily overlapping bounding boxes.

Table 2: Comparison between the joint-based and the HMR [19]-style regressors on 3DPW-Crowd. We used 2D pose outputs from Cheng *et al*. [6]. The joint-based regressor outperforms the HMR-style regressor on crowd scenes.

parameter regressor type	MPJPE	PA-MPJPE
HMR [19]-style regressor	89.0	59.5
joint-based regressor (Ours)	**85.6**	**55.4**

Table 3: Ablation on 3DPW-Crowd about 3DPoseNet that recovers a 3D pose from the 2D human pose-guided image feature. We used 2D pose outputs from Cheng *et al*. [6]. The results validate that 3DPoseNet increases the robustness of our system on crowd scenes.

network	MPJPE	PA-MPJPE
without 3DPoseNet	96.7	60.1
2DPoseNet	88.3	56.4
3DPoseNet (Ours)	**85.6**	**55.4**

Joint-based regressor vs. HMR-style regressor. Table 2 shows that the joint-based regressor outperforms the HMR [19]-style regressor on 3DPW-Crowd. The results prove that the joint-based regressor better preserves the 2D pose cues for the target person in crowd scenes. Note that the joint-based regressor utilizes joint-level feature sampled from (x,y) positions of the predicted 3D pose. In contrast, the HMR-style regressor regresses SMPL parameters from the global-average-pooled feature.

Validity of 3DPoseNet. Table 3 validates 3DPoseNet by comparing it with its variations. The first-row network samples the joint-level feature based on 2D joint predictions from off-the-shelf 2D pose estimators. The second-row network, 2DPoseNet, estimates a 2D pose instead of a 3D pose. Both networks concatenate the joint-level feature and 2D joint coordinates, and feed it to 3DShapeNet.

The accuracy significantly drops without 3DPoseNet on 3DPW-Crowd. The result proves that the refining process of the 2D pose input is essential for the joint-based regressor of 3DCrowdNet. We can also verify the effectiveness of estimating a 3D pose in Table 3. The clear accuracy improvement proves that the depth information can be reliably estimated from a 2D pose and image feature, and it is bene-

Table 4: Comparison between 3DCrowdNet and previous methods on 3DPW-Crowd. $*$ means that we used CrowdPose [23] train data.

method	MPJPE	PA-MPJPE	MPVPE
SPIN [21]	121.2	69.9	144.1
Pose2Mesh [7]	124.8	79.8	149.5
I2L-MeshNet [32]	115.7	73.5	162.0
3DCrowdNet (Ours)	86.9	56.8	110.3
3DCrowdNet (Ours)*	**85.6**	**55.4**	**109.0**

Table 5: Comparison between 3DCrowdNet and previous methods on MuPoTS [29]. The numbers are 3DPCK for all annotations (All) and annotations matched to a prediction (Matched). The tiny scripts indicate the source of bounding boxes or input 2D poses of our system.

method	All	Matched
SMPLify-X [38] (Cao *et al.* [2])	62.8	68.0
HMR [19] (Cao *et al.* [2])	66.0	70.9
HMR [19] (He *et al.* [11])	65.6	68.6
Jiang *et al.* [16]	69.1	72.2
3DCrowdNet (Ours) (Cao *et al.* [2])	70.2	70.9
3DCrowdNet (Ours) (Cheng *et al.* [6])	**72.7**	**73.3**

ficial for the accuracy of the final output.

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

Unless indicated, our 3DCrowdNet is not trained on CrowdPose [23] train set in this section.

3DPW-Crowd. We compare our 3DCrowdNet with [7, 21, 32] on 3DPW-Crowd in Table 4. They are recent state-of-the-art 3D human pose and shape estimation methods on 3DPW [49], and publicly released the codes for an evaluation. Our approach outperforms SPIN [56], which takes only the image feature as input and uses a HMR [19]-style regressor. The result is coherent with the results in Table 1 and 2 of our ablation studies. 3DCrowdNet also defeats Pose2Mesh, a method that can benefit from crowd-scene robust 2D pose outputs. We used the same 2D pose outputs from Cheng *et al.* [6] for Pose2Mesh and 3DCrowdNet. While Pose2Mesh heavily depends on the accuracy of the 2D pose outputs, our system improves the accuracy of pose and shape using rich depth and shape cues in the image feature. Figure 6 supports our statement. Leveraging the image feature, 3DCrowdNet reconstructs a 3D shape that best describes the target person in images, even when the 2D pose is inaccurate.

MuPoTS. Table 5 compares our 3DCrowdNet with [16, 19, 38]. As the second and fifth rows show, 3DCrowdNet outperforms HMR [2] when both methods are based on the 2D poses from Cao *et al.* [2]. While HMR [2] utilizes the 2D pose only to get a bounding box, 3DCrowdNet additionally uses the 2D pose to guide a network to focus on a target person from crowd scenes. Leveraging more information in given outputs is natural, and leads to better accuracy. Moreover, our 3DCrowdNet benefits from improved 2D pose outputs from Cheng *et al.* [6]. In summary, Table 5 proves that 3DCrowdNet provides accurate 3D human pose and shape in images with multi-person, and has the potential to improve with better 2D pose outputs. The numbers of other methods in Table 5 are from Jiang *et al.* [16].

Table 6: Comparison between 3DCrowdNet and previous methods on CMU-Panoptic [18]. We follow the publicly released evaluation protocol of Jiang *et al.* [16].

method	Haggl.	Mafia	Ultim.	Pizza	Mean
Zanfir *et al.* [54]	140.0	165.9	150.7	156.0	153.4
Zanfir *et al.* [55]	141.4	152.3	145.0	162.5	150.3
Jiang *et al.* [16]	129.6	**133.5**	153.0	156.7	143.2
3DCrowdNet (Ours)	**109.60**	135.9	**129.8**	**135.6**	**127.6**

Table 7: Comparison on 3DPW [49] between 3DCrowdNet and previous methods of 3D human pose and shape estimation from a single image. 3DCrowdNet achieves the best accuracy in all metrics. All methods except Zanfir *et al.* [53] did not use 3DPW train data.

method	MPJPE	PA-MPJPE	MPVPE
HMR [19]	130	76.7	-
GraphCMR [22]	-	70.2	-
SPIN [21]	96.9	59.2	116.4
STRAPS [45]	-	66.8	-
Zanfir *et al.* [53]	90.0	57.1	-
Pose2Mesh [7]	88.9	58.3	106.3
I2L-MeshNet	93.2	57.7	110.1
Song *et al.* [46]	-	55.9	-
3DCrowdNet (Ours)	**82.8**	**52.2**	**100.2**

CMU-Panoptic. Table 6 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 crowd scenes with daily social activities. Note that no data from CMU-Panoptic are used for training.

3DPW. Table 7 shows that 3DCrowdNet achieves the state-of-the-art accuracy on general in-the-wild scenes. The result validates that 3DCrowdNet is robust to diverse challenges in in-the-wild scenes, although our method is designed to target crowd scenes. The left samples in the second and third rows of Figure 5 depicts the robustness of 3DCrowdNet to truncation and object occlusion in images, which are common in general in-the-wild scenes. The proposed system reconstructs plausible full 3D pose and shape from image feature and partial 2D poses of truncated and

Figure 5: Qualitative comparison on the CrowdPose [23] test set. From left, an input image, 3DCrowdNet outputs, SPIN [21] outputs. Both methods used the same bounding boxes computed from 2D pose outputs of Cheng *et al*. [6].

Figure 6: Visualization of 3D shapes from different viewpoints. The proposed 3DCrowdNet can recover 3D pose and shape that best describes the target person in an image, even when provided with incorrect 2D pose outputs, using both the 2D pose output and image feature.

occluded bodies.

We provide the qualitative comparison between 3DCrowdNet and SPIN [21] in Figure 5. SPIN [21] is the most relevant competitor among current state-of-the-art methods, since it leverages image feature and estimates SMPL parameters as 3DCrowdNet. Apparently, 3DCrowdNet produces much more robust 3D shapes on crowd scenes than SPIN. 3DCrowdNet disentangles different people from a target person in a bounding box, and estimates reasonable 3D pose and shape under diverse occlusion, including inter-person occlusion. We provide more qualitative comparison with other methods [7, 32] and failure cases in the supplementary material.

6. Discussion and future works

Our approach is the first to provide robust 3D human pose and shape outputs of each individual in in-the-wild crowd scenes. However, it is still demanding to recover accurate 3D poses from images with extremely close interactions, such as tussles in sports. The extreme cases often involve challenging poses and severe inter-person occlusion, both of which are rare in existing train data. Although [23, 41, 57] recently proposed crowd datasets, they only provide 2D annotations and a limited number of images. As a result, the effect of additionally training on such datasets is marginal, as shown in the fourth and fifth rows of Table 4. Future works should suggest a 3D train dataset that covers diverse poses and occlusion in crowd scenes.

7. Conclusion

We present 3DCrowdNet, the first single image-based 3D human pose and shape estimation system that targets in-the-wild crowd scenes. Previous methods suffer from noisy human bounding boxes and inter-person occlusion in crowd scenes. We resolve the issue by leveraging crowd-scene robust 2D pose outputs. The proposed 3DCrowdNet is capable of focusing on a target person in crowd scenes, and effectively utilizes the articulation information in the 2D poses. We provide extensive ablation studies for the crowd-scene robust system, which would give meaningful insight for future works. Also, we show that 3DCrowdNet provides far more robust outputs on in-the-wild crowd scenes than previous methods numerically and qualitatively.

Supplementary Material *for*

3DCrowdNet: 2D Human Pose-Guided 3D Crowd Human Pose and Shape Estimation in the Wild

Figure 7: Visualization of 3D shapes from different viewpoints, given inaccurate input 2D poses. Our 3DCrowdNet can recover a 3D shape that best describes the target person in an image, even when provided with inaccurate 2D poses, using both the 2D pose and image feature.

8. More qualitative results.

Pose correction. Figure 7 shows that our 3DCrowdNet can estimate robust 3D pose and shape given inaccurate 2D poses in crowd scenes. Due to inter-person occlusion and overlapping bounding boxes between people, 2D pose estimators [2, 6] may produce inaccurate joint predictions as shown in the first, second, and third rows. In this regard, 3DCrowdNet assigns don't-care values to the joint predictions with low confidence (*e.g.* lower than 0.1 for outputs from Cheng *et al*. [6]) at first. Then 3DPoseNet of 3DCrowdNet refines the inaccurate input 2D pose using image feature that contains the context information in images. Last, 3DShapeNet of 3DCrowdNet estimates human model parameters, SMPL [26] parameters. The whole process makes our system predict a 3D human shape that best describes a target person in images and is plausible, even when provided with inaccurate 2D joint predictions. The fourth and fifth rows prove that our approach is also effective on estimating robust 3D shapes for truncated images.

Comparison with I2L-MeshNet [32] and Pose2Mesh [7]. Figure 8 shows the qualitative comparison between 3DCrowdNet, I2L-MeshNet [32], and Pose2Mesh [7]. Among the three methods, 3DCrowdNet apparently produces much more robust 3D shapes on the crowd scenes. I2L-MeshNet tends to provide noisy body shape estimation in crowd scenes. It misses a person in images with overlapping bounding boxes as depicted in the fifth and sixth rows. Limb outputs in the second, fourth, and sixth rows are inaccurately estimated or swapped with other people. Overall, I2L-MeshNet is hard to disentangle people in overlapping bounding boxes and provides inaccurate outputs when a body is occluded in crowd scenes.

Pose2Mesh estimates more stable 3D shapes than I2L-MeshNet in crowd scenes, but the 3D poses of Pose2Mesh are different from the poses depicted in the input images. The legs of people in the second, third, sixth rows are inaccurately estimated. Especially, Pose2Mesh wrongly predicts the legs of the most right person (apricot color) in the third row; although the input 2D pose is correct. Also, in the first row, the most front person's global orientation prediction (pink color) is inaccurate. The evidence proves that solely relying on 2D geometry information and not leveraging image feature has limitations for crowd-scene robust 3D shape estimation, since the approach produces plausible 3D shapes for given 2D geometry, not 3D shapes that best describe a person in images.

To sum up, our 3DCrowdNet estimates the most accurate 3D pose and shape from in-the-wild crowd scenes compared to the competitors [7, 32] by leveraging image feature, the guidance of crowd-scene robust 2D pose outputs, and the articulation information in 2D poses.

Failure cases of 3DCrowdNet. As Figure 9 shows, 3DCrowdNet falls short in images with extreme interactions, such as tussles in sports. Input 2D poses, which are the outputs of Cheng *et al*. [6], are highly inaccurate in such crowd scenes, and 3DCrowdNet does not effectively estimate the correct 3D pose and shape from them. Severe inter-person occlusion, challenging poses, and similar ap-

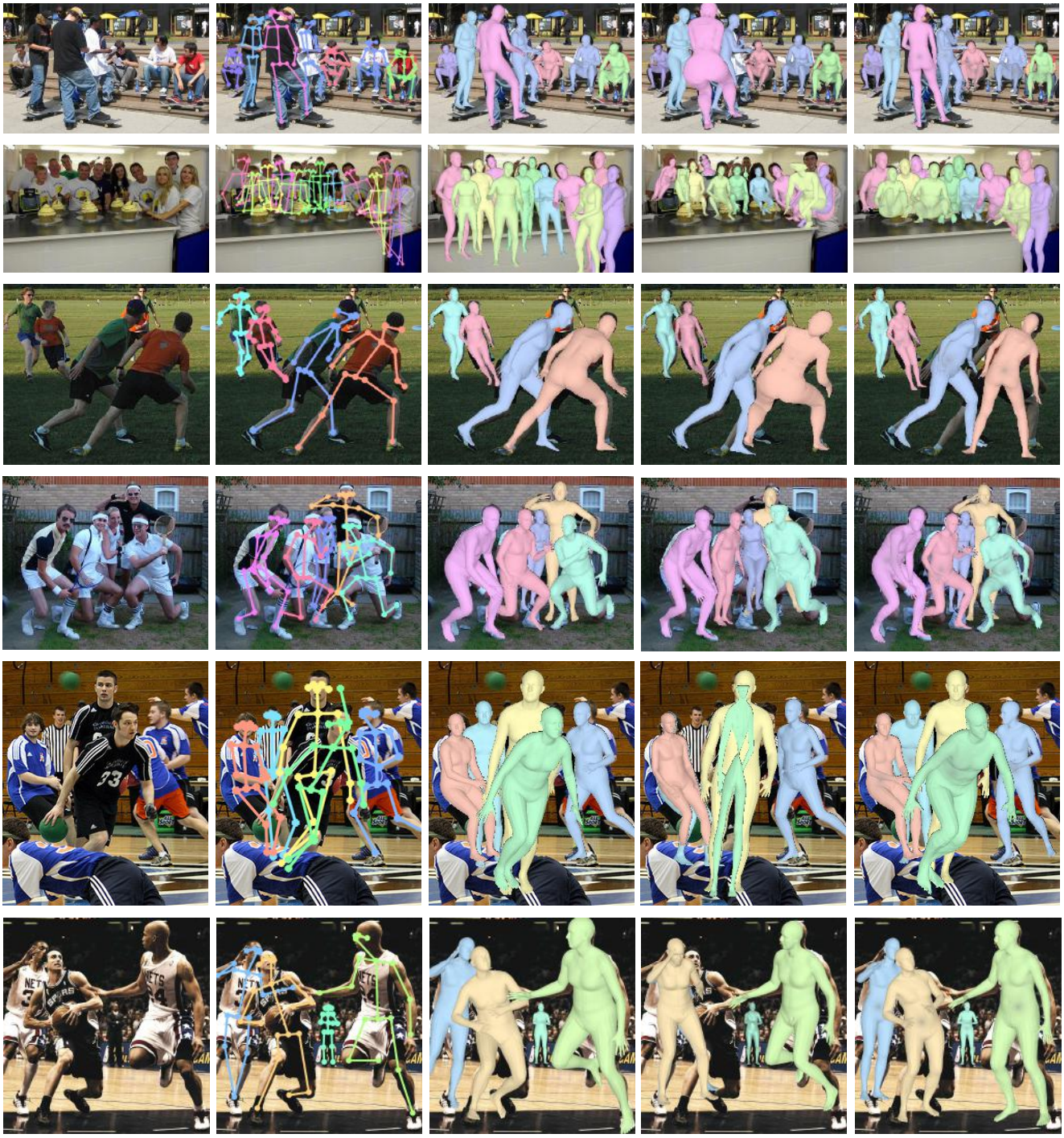


Figure 8: Qualitative comparison on the CrowdPose [23] test set. From left, an input image, input 2D poses estimated by Cheng *et al*. [6], 3DCrowdNet, I2L-MeshNet [32], and Pose2Mesh [7] outputs. Our 3DCrowdNet successfully disentangles a target person from other people in a bounding box compared to I2L-MeshNet. Also, 3DCrowdNet produces a 3D shape that best describes a target person in images, while Pose2Mesh estimates a plausible 3D shape for given 2D poses, which could not correspond to images. 3DCrowdNet and I2L-MeshNet use the same bounding boxes to crop an image for each person. 3DCrowdNet and Pose2Mesh use the same 2D poses from Cheng *et al*. [6].

Figure 9: Failure cases of 3DCrowdNet.

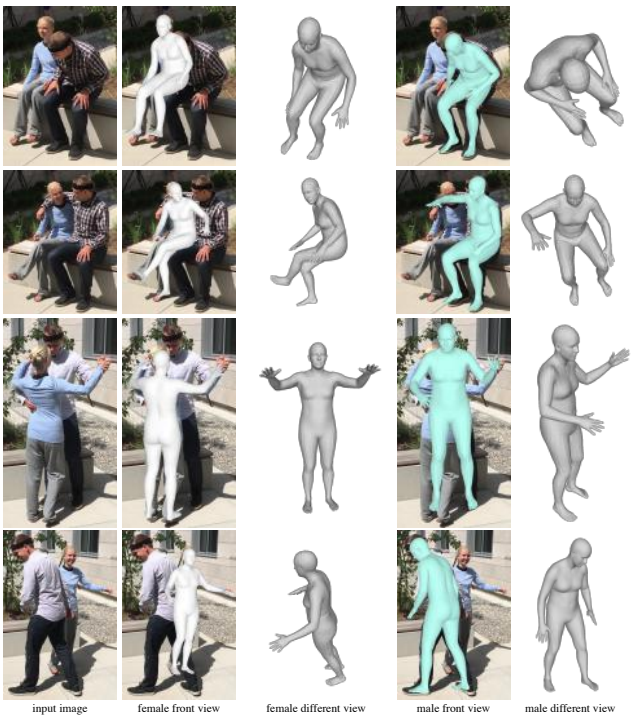


Figure 10: 3DCrowdNet outputs on 3DPW-Crowd.

pearance of clothes are the main causes of the failures. For example, in the left bottom image, 3DCrowdNet wrongly predicts a green-colored person's right leg, who is tackling in a soccer game. Data with the above situations is scarce in existing train datasets, and a large-scale train dataset that primarily covers such crowd scenes should be introduced.

Results on 3DPW-Crowd. Figure 10 illustrates the 3DCrowdNet results on 3DPW-Crowd. 3DCrowdNet estimates robust 3D pose and shape on images that show people having highly close interaction. Different people in overlapping bounding boxes are disentangled, and occluded body parts are reasonably reconstructed.

9. Datasets.

3DPW-Crowd. The sequence names of 3DPW-Crowd are *courtyard_hug_00* and *courtyard_dancing_00*, a subset of the 3DPW [49] validation set. 3DPW-Crowd contains 1073 images and 1923 persons with GT 3D pose and shape annotations. The average bounding box IoU is 37.5%, and the CrowdIndex [23] is 49.3%. We used 14 joints defined by Human3.6M [15] for evaluating PA-MPJPE and MPJPE following the previous works [7, 21, 32].

Figure 11 presents sampled frames of 3DPW-Crowd sequences in order. The sequences reveal multiple challenges of in-the-wild crowd scenes, such as heavy inter-person occlusion and overlapping bounding boxes.

MuPoTS. MuPoTS [29] contains 20 sequences, 8370 images, and 20899 persons with GT 3D pose annotations. The sequences are captured indoors and outdoors, and GT 3D poses are obtained by a multi-view marker-less motion capture system. The average bounding box IoU is 3.8%, and the CrowdIndex [23] is 13.2%. We used the official MATLAB code for evaluation.

CMU-Panoptic. We selected four sequences that show people doing social activities, namely *Haggling*, *Mafia*, *Ultimatum*, and *Pizza* following [16, 54, 55]. Sequences captured by the 16th and 30th cameras are selected. The sequences contain 9600 frames and 21,404 persons with GT 3D pose annotations. The average bounding box IoU is 2.0%, and the CrowdIndex [23] is 11.1%. We used preprocessed GT annotations and followed the evaluation protocol of Jiang *et al*. [16] in their official code repository.

courtyard_hug_00

courtyard_dancing_00

Figure 11: Sampled frames of 3DPW-Crowd sequences in order.

3DPW. We used the test set of 3DPW [49] following the official split protocol. The test set contains 26240 images and 35515 persons with GT 3D pose and shape annotations. The average bounding box IoU is 3.7%, and the CrowdIndex [23] is 4.9%. Sequences starring one actor are excluded in computing the bounding box IoU and the CrowdIndex. We used 14 joints defined by Human3.6M [15] for evaluating PA-MPJPE and MPJPE following the previous works [7, 21, 32].

10. 2D pose estimators.

In this work, we used 2D pose outputs from Cao *et al*. [2] and Cheng *et al*. [6]. The outputs from Cao *et al*. [2] used in 3DPW-Crowd and 3DPW [49] are included in the annotations of 3DPW [49]. The outputs from Cao *et al*. [2] used in MuPoTS [29] are obtained by running the third-party PyTorch [37] code implementation. The model is trained on **COCO2017 train** [24] dataset and achieves 0.653 mAP (mean Average Precision) on **COCO2017 val** dataset. The outputs from Cheng *et al*. [6] are all obtained by running the official code implementation. The model is trained on **COCO2017 train** dataset and achieves 0.671 mAP on **COCO2017 val** dataset.

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. 2, 3, 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, 4, 6, 7, 9, 12

[3] Ju Yong Chang, Gyeongsik Moon, and Kyoung Mu Lee. Absposelifter: Absolute 3d human pose lifting network from a single noisy 2d human pose. *arXiv preprint arXiv:1910.12029*, 2020. 4

[4] He Chen, Pengfei Guo, Pengfei Li, Gim Hee Lee, and Gregory Chirikjian. Multi-person 3d pose estimation in crowded scenes based on multi-view geometry. In *ECCV*, 2020. 3

[5] 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

[6] 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, 4, 6, 7, 8, 9, 10, 12

[7] 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. 3, 4, 5, 7, 8, 9, 10, 11, 12

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *CVPR*, 2017. 3

[9] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 1, 3

[10] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, 2019. 3, 4

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 3, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 5

[13] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 3

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014. 2, 3, 5, 11, 12

[16] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 3, 5, 7, 11

[17] 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

[18] 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. *TPAMI*, 2017. 2, 5, 7

[19] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 3, 5, 6, 7

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

[21] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1, 3, 5, 7, 8, 11, 12

[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 3, 7

[23] 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. 3, 5, 7, 8, 10, 11, 12

[24] 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, 12

[25] 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. 4

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 1, 9

[27] 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, 5

[28] 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

[29] 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, 11, 12

[30] 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

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

[32] 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. 3, 5, 7, 8, 9, 10, 11, 12

[33] 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

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

[35] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural Body Fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 3

[36] 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

[37] 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, 12

[38] 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. 7

[39] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 1, 3

[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. 3, 8

[42] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[43] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 3

[44] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *TPAMI*, 2019. 3

[45] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *BMVC*, 2020. 3, 7

[46] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 3, 7

[47] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 4, 5

[48] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 3

[49] 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, 7, 11, 12

[50] 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

[51] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 2

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

[53] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *ECCV*, 2020. 3, 5, 7

[54] 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, 11

[55] 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. 3, 5, 7, 11

[56] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3d human shape and pose from dense body parts. *TPAMI*, 2020. 3, 4, 7

[57] 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, 8

[58] 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