Putting People in their Place: Monocular Regression of 3D People in Depth **Supplementary Material**

Figure 1. More qualitative results on Internet images [1].

1. Introduction

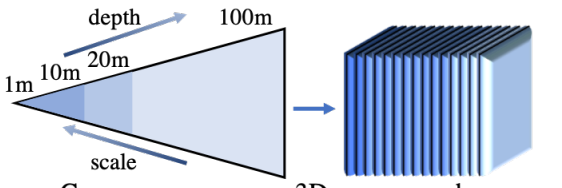
In this material, we provide more implementation details, analysis of the "*Relative Human*" (RH) dataset, and quantitative/qualitative comparisons to the state-of-the-art methods. Additionally, we present more visual results, like Fig. 1, to show the performance of BEV under different situations and to explore its failure modes.

2. Implementation Details

In this section, we introduce the details of our camera representation, network architecture, and training details.

2.1. Normalized Camera Representation

To supervise 3D joints $\vec{J}$ with 2D poses, existing methods [13, 30] widely adopt a weak-perspective camera model



Depth range(m)	0-1	1-10	10-20	20-100
Percentage of maps	2/32	25/32	3/32	2/32

Figure 2. Pre-defined 3D camera anchor maps.

to project $\vec{J}$ onto the image plane. For better depth reasoning, we employ a perspective camera model to perform this 2D projection.

In most cases, accurate camera parameters for in-the-wild images are unavailable. In this situation, to avoid reliance on the camera parameters of 2D projection, we assume that the input image is captured with a standard camera without radial distortion. Then we can assign static values for the field of view (FOV) and image size $\vec{W}$ of this standard camera. The focal length $\vec{f} = (f_x, f_y)$ can be defined as $\vec{W}/(2tan(FOV/2))$. Given the 3D translation (x_i, y_i, d_i) of i-th subject and the focal length, the 2D projection $(\vec{u_i}, \vec{v_i})$ of 3D joints $(\vec{J_i^x}, \vec{J_i^y}, \vec{J_i^d})$ is defined as

$$\vec{u_i} = \frac{f_x(\vec{J_i^x} + x_i)}{\vec{J_i^d} + d_i}, \vec{v_i} = \frac{f_y(\vec{J_i^y} + y_i)}{\vec{J_i^d} + d_i}. \quad (1)$$

In cases where the camera parameters are provided, we can convert the 3D translation estimated in our standard camera space to the given one. With K pairs of estimated 3D joints $\vec{J}$ and their 2D projection (obtained via Eq. 1), we can solve the 3D translation at a specific camera space via a PnP algorithm (e.g. RANSAC [7]).

However, in the image, 3D translation is not as intuitive as the person's scale used by weak-perspective methods. For instance, a small 2D scale change in an image may correspond to a large difference in 3D translation in camera space, especially for people who are far away in depth. Therefore, to alleviate this difference, we convert the 3D translation (x_i, y_i, d_i) to a normalized scale-based format (s_i, t_i^y, t_i^x) via a scale factor $s_i = (d_i tan(FOV/2))^{-1}$, where $t_i^y = y_i s_i, t_i^x = x_i s_i$. The normalized representation is proportional to the person's scale. When FOV=60°, the sensitive part $s_i \in (0, 2)$ corresponds to $d_i \in (0.86, +\infty)$ in meters, which is more suitable for the network to estimate.

Additionally, we observe that people in the depth range (1m,10m) show more abundant and stable information in pose, shape, and depth, which deserve more attention. Additionally, most of our training samples are within this depth range. As we introduced in the main paper, 3D camera anchor maps define the way we voxelize the 3D camera space. Therefore, we adjust the occupancy ratio of different depths in the channel number of 3D camera anchor maps. As shown in Fig. 2, we first split the camera space into 4 regions in depth and then evenly put the different number (shown in the table) of 3D camera anchor maps inside each region. For instance, we put 25/32 3D camera anchor maps inside the depth range (1m,10m); this gives more attention to this critical depth range. Each anchor map contains the normalized camera values (s_i, t_i^y, t_i^x) at the corresponding position.

2.2. Network Architecture

We develop a bird's-eye-view-based coarse-to-fine localization pipeline to estimate the 3D translation of all people in the scene in one shot. In Fig. 3, we present the network architecture of estimating five 2D maps and two 3D maps, which are used to generate the final results as shown in Fig. 2 of the main paper. The input size $\vec{W}$ is $(512, 512)$. Following ROMP [30], we adopt a multi-head architecture and use HRNet-32 [3] as backbone. With backbone feature maps of size $\mathbb{R}^{32\times H\times W}$, we employ three head branches to estimate four front-/bird's-eye-view 2D maps and a Mesh Feature map.

As illustrated in Fig. 4, our key design is to convert the front-view features to a bird's eye view via explicit operations including height-wise suppression and depth-wise exploration. As shown in the middle branch of Fig. 3, we first explore the depth information of backbone features via a Bottleneck block. And then we concatenate the explored depth features and front-view 2D maps as input to the BVH branch. As shown in Fig. 4, we compress the 2D feature maps in height to obtain 1D feature vectors. In the BVH branch (Fig. 3), we employ six 1D convolution blocks to explicitly explore features in depth. Two bird's-eye-view maps are of size $\mathbb{R}^{1\times D\times W}$.

Next, we compose the front-view and bird's-eye-view maps to generate 3D maps. We extend the front-view maps with an additional depth dimension and repeat D times. We also extend the bird's-eye-view maps with an additional height dimension and repeat H times. To obtain the 3D Center map, we multiply the bird's-eye-view Body Center heatmap to the front-view one and refine it with a 3D refiner (Fig. 3). Then we add the bird's-eye-view offset map to the last channel of the front-view one to refine the depth. To obtain the 3D Offset map, we further use a 3D refiner to refine the composed 3D maps, which improves the consistency between features of two views.

2.3. Datasets

In this section, we introduce the datasets we used during training and evaluation.

AGORA [27] is a synthetic dataset with accurate annota-

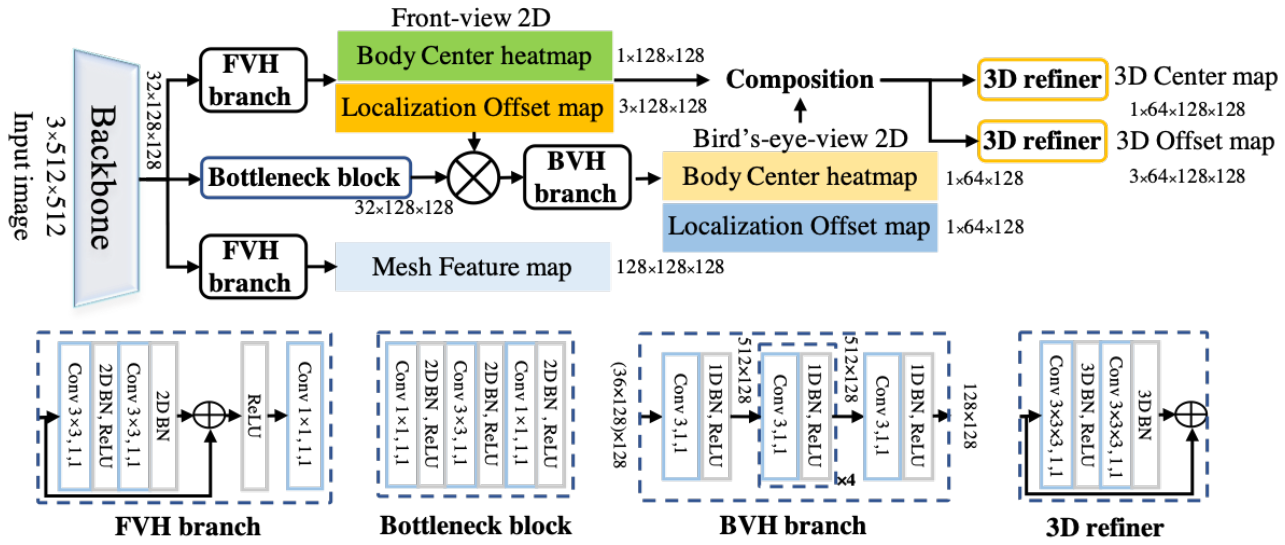


Figure 3. Network architecture.

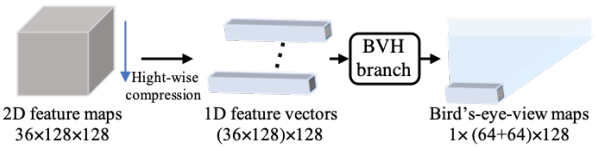


Figure 4. Operations to convert the front-view features to a bird's eye view (shown in 3D camera space represented by 3D camera anchor maps).

tions of body meshes and 3D translations, with 4,240 high-realism textured scans in diverse poses and clothes. Importantly, it contains 257 child scans. It contains 14K training and 3K test images. Each image has 5-15 people with frequent occlusions. **AGORA-PC [27]** is a high occlusion subset of the AGORA validation set. Each image has over 70% occlusion. We use it to evaluate the performance under severe occlusion. Note that there are no child samples in the validation set.

Human3.6M [8] is a single-person 3D pose dataset. It contains videos of 9 professional actors performing activities in 17 scenarios. It provides 3D pose annotations for each frame. We sample every 5 frames to reduce redundancy. We use its training set for training.

MuCo-3DHP [23] is a synthetic multi-person 3D pose dataset. It is built on the single-person 3D pose dataset, MPI-INF-3DHP [23]. They use segmentation annotations to blend multiple single-person images into one. For a fair comparison with 3DMPPE [25], we use the same synthetic version for training.

Other 2D pose datasets. For better generalization, we also use four 2D pose datasets for training, including COCO [21], MPII [2], LSP [10], and CrowdPose [18]. Besides, we also use the pseudo-3D annotations [12] for training.

2.4. Training Details

The size of output maps are $H = W = 128$, and $D = 64$. The threshold for the age offset is set to $t_\alpha = 0.8$. The FOV is set to 60°. The loss weights are $w_{mpj} = 200$, $w_{pmpj} = 360$, $w_{pj2d} = 400$, $w_\theta = 80$, $w_\beta = 60$, $w_{prior} = 1.6$, $w_{cm} = 100$, $w_{cm3d} = 1000$, $w_{age} = 4000$, and $w_{depth} = 400$. We train BEV on a server with four Tesla V100 GPUs. The batch size is 64. The learning rate is $5e^{-5}$. The confidence threshold of the Body Center heatmap is 0.12.

Additionally, although we strive to alleviate the age bias in training samples, the age bias in existing 3D pose datasets is severe, and we have to use them to obtain good 3D pose estimation. To handle the imbalanced distribution of the training sample space, we balance the sampling ratio of different datasets and evenly select the training samples from different age groups on RH. The sampling ratios of different datasets are 16% AGORA, 16% MuCo-3DHP, 16% RH, 18% Human3.6M, 14% COCO, 8% CrowdPose, 6% MPII, and 6% LSP.

Also, we adopt a two-step training strategy. We first learn monocular 3D pose and shape estimation for 120 epochs on basic training datasets. Then we add the weak annotations of RH to training samples and train for 120 epochs. If we need to fine-tune on AGORA, we add AGORA to the training sequence and train for 80 epochs. In this process, the validation set of the RH is used to select checkpoints with good performance.

2.5. Processing High-resolution Images

As a one-stage method, BEV takes an image of constant size as input. However, to process the high-resolution images, directly resizing them to a constant size would sacrifice the performance. Therefore, we develop a sliding window-based pipeline to achieve promising results on high-resolution images, as shown in Fig. 1 of the main paper. In detail, we evenly split the image into multiple grids and then apply BEV on each grid. This process is similar to the sliding window operation of 2D convolution. At each grid, we only take the result whose body center falls in the center area of the grid. Then we perform non-maximum suppression on the edge between grids to get rid of redundant predictions. In this process, overlapping predictions with lower center confidence values will be deleted.

3. Relative Human Dataset

In this section, we provide more detailed analyses of our Relative Human dataset.

In total, we collect about 7,689 images with weak annotations of 24,814 people. We split them into three groups (5218, 635, 1836) for training, validation, and test respectively. Among these images, about 1,000 images are collected from a free photo website [1] and we annotate the 2D poses defined as Fig. 5. Note that compared with LSP's 14 keypoints, we add keypoints on the face and feet to represent their orientations. The remaining images are selected from existing 2D pose datasets [18,21,33]. We correct some erroneous 2D poses from the existing 2D pose dataset and add the missing detections. Note that a large number of images in CrodPose [18] and OCHuman [33] are selected from COCO [21] and MPII [2], which are also used as training samples by our compared methods [9,15,16,30]. Therefore, we use these common images for training.

We classify all people in the image into four age groups, baby, child, teenager, and adult according to the following age ranges: baby (0-3), kid (3-8), teenager (8-16), and adult (16+). As shown in Tab. 1, we provide the number of subjects in the four age groups and their proportions. Compared with the existing multi-person 3D pose datasets [23, 27, 31], RH contains richer subjects and more occlusion cases. Therefore, RH is more general and suitable for evaluating depth reasoning in the wild.

The consistency of weak annotations. During the collection of weak annotations, we observe that people's judgments for such weak labels vary greatly. It is hard to obtain consistent weak labels through online platforms (e.g. AMT). Therefore, offline, we organized a group of labelers and trained them with unified standards. To test how well they learn the standards, we prepare some pre-labeled data as test samples. Ones who pass the test after training were employed for official labeling. In addition, the annotations are double-checked by professional testers and the author.

RH splits	Babies	Children	Teenagers	Adults
All	1534 / 6%	2720 / 10%	1067 / 4%	19493 / 78%
Train	942 / 5%	1795 / 10%	690 / 4%	13478 / 79%
Validation	117 / 5%	209 / 9%	101 / 4%	1680 / 79%
Test	475 / 8%	716 / 12%	276 / 4%	4335 / 74%
Existing multi-person 3D pose datasets				
MuPoTS [23]	-	-	-	8 / 100%
3DPW [31]	-	-	-	18 / 100%
AGORA [27]	-	257 / 6%	-	3983 / 94%

Table 1. Subject number/proportions of four age groups on Relative Human (RH) and 3D pose benchmarks.

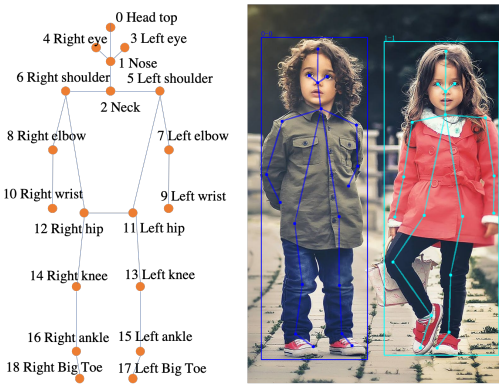


Figure 5. The 2D skeleton definition.

4. Discussion

Why not estimate the 3D heatmap directly? The main challenge is the lack of sufficient multi-person data with accurate 3D translation annotations for supervision, especially for in-the-wild cases. Due to the data lack problem, directly learning 3D heatmap performs poorly. It is hard to effectively supervise multi-person 3D heatmaps with 2D annotations. In contrast, our separable representation disentangles the 3D heatmap into the front-view and the bird's-eye-view maps. In this way, our model can learn robust front-view localization from abundant 2D in-the-wild datasets. With robust front-view attention, the model can focus on learning depth reasoning from weak annotations in RH with the proposed WST.

Heatmap refinement and decomposition. Following previous methods [30], we also adopt the powerful heatmap representation for detection. While its rough granularity limits its effectiveness in fine localization. Some previous methods explore the refinement and decomposition of the heatmap to alleviate this problem. PifPaf [17] estimates offset maps to refine the coarse 2D pose coordinates parsed from the heatmap. VNect [24] estimates three 2D maps

Input		Method	F1 score↑	Precision↑	Recall↑	MVE↓	MPJPE↓	NMVE↓	NMJE↓
Multi-stage	**Person crops from 3840x2160**	HMR [13]	0.38	0.27	0.61	209.3	219.4	550.8	577.4
		SMPLify-X‡ [28]	**0.57**	**0.60**	0.55	213.3	208.3	374.2	365.4
		EFT [12]	0.43	0.34	0.60	193.5	202.7	450.0	471.4
		SPIN [16]	0.33	0.23	0.61	193.2	203.7	585.5	617.3
		ExPose [5]	0.53	0.46	0.61	174.0	176.6	328.3	333.2
		Frankmocap [29]	0.40	0.30	0.62	204.2	203.7	510.5	509.2
		PyMAF [32]	0.27	0.16	0.82	192.0	203.2	711.1	752.6
		PIXIE [6]	0.48	0.39	0.61	174.6	174.7	363.8	364.0
		SPIN⋆ [27]	0.31	0.21	0.60	186.7	191.7	602.3	618.4
		SPEC⋆ [15]	0.52	0.40	0.73	163.2	171.0	313.8	328.8
		PARE [14]	0.55	0.44	0.74	186.4	193.9	338.9	352.5
		Pose2Pose⋆† [26]	0.56	0.40	**0.91**	**146.4**	**153.3**	**261.4**	**273.8**
One-stage	**512x512**	ROMP [30]	0.38	0.39	0.37	198.5	207.4	522.4	545.8
		BEV w/o WST	0.41	0.39	0.45	194.4	202.6	474.1	494.1
		ROMP⋆ [30]	0.50	0.37	0.80	156.6	159.8	313.2	319.6
		BEV⋆ w/o WST	**0.58**	**0.44**	**0.86**	146.0	148.3	251.7	255.7
		BEV⋆	0.55	0.41	0.85	**125.9**	**129.1**	**228.9**	**234.7**

Table 2. Comparison to existing SOTA methods on the "AGORA kids" test set. Results are obtained from the AGORA leaderboard. ⋆ is fine-tuning on the AGORA training set or synthetic data [15] generated in the same way as AGORA. ‡ means the optimization-based method while the rest are learning-based methods. † means the paper is under review.

Input		Method	F1 score↑	Precision↑	Recall↑	MVE↓	MPJPE↓	NMVE↓	NMJE↓
Multi-stage	**Person crops from 3840x2160**	HMR [13]	0.80	0.93	0.70	173.6	180.5	217.0	226.0
		SMPLify-X ‡ [28]	0.71	0.86	0.60	187.0	182.1	263.3	256.5
		EFT [12]	0.69	**0.97**	0.54	159.0	165.4	196.3	203.6
		SPIN [16]	0.78	0.91	0.69	168.7	175.1	216.3	223.1
		ExPose [5]	0.82	0.96	0.71	151.5	150.4	184.8	183.4
		Frankmocap [29]	0.80	0.93	0.71	204.2	203.7	510.5	509.2
		PyMAF [32]	0.84	0.86	0.82	192.0	203.2	711.1	752.6
		PIXIE [6]	0.82	0.95	0.73	142.2	140.3	173.4	171.1
		SPIN⋆ [27]	0.77	0.91	0.67	168.7	175.1	216.3	223.1
		SPEC⋆ [15]	0.84	0.96	0.74	106.5	112.3	126.8	133.7
		PARE [14]	0.84	0.96	0.75	140.9	146.2	167.7	174.0
		Pose2Pose⋆† [26]	**0.94**	0.94	**0.93**	**84.8**	**89.8**	**90.2**	**95.5**
One-stage	**512x512**	ROMP [30]	0.69	0.97	0.54	161.4	168.1	233.9	242.3
		BEV w/o WST	0.75	**0.97**	0.61	164.2	169.1	218.9	225.5
		ROMP⋆ [30]	0.91	0.95	0.88	103.4	108.1	113.6	118.8
		BEV⋆ w/o WST	0.93	0.96	0.90	105.6	109.7	113.5	118.0
		BEV⋆	**0.93**	0.96	**0.90**	**100.7**	**105.3**	**108.3**	**113.2**

Table 3. Comparison to existing SOTA methods on AGORA full test set. Results are obtained from the AGORA leaderboard. ⋆ is fine-tuning on the AGORA training set or synthetic data [15] generated in the same way as AGORA. ‡ means the optimization-based method while the rest are learning-based methods. † means the paper is under review.

containing x/y/z coordinates of the 3D pose at each position. Luvizon et al. [22] employ soft-argmax to decompose 2D/3D heatmap into 1D for separate supervision, while it does not deal with multiple overlapping people. Different from previous solutions, we propose a novel bird's-eye-view-based representation for multi-person 3D localization. As we introduced above, it disentangles the depth-wise information into an individual map for easier learning. We also estimate a 3D offset map to improve the granularity of 3D localization.

5. Quantitative and Qualitative Results

In this section, we first show more comparisons to SOTA methods on AGORA and then provide more qualitative results on Internet images, CMU Panpotic [11], AGORA [27], and RH.

5.1. Quantitative Comparisons

In Tab. 2 and 3, we show the results of existing SOTA methods on "AGORA kids" and the full test set respectively. Results in Tab. 2 show that BEV outperforms all previous methods by a large margin in terms of child mesh

Figure 6. Qualitative results on CMU Panoptic [11] and AGORA [27] datasets.

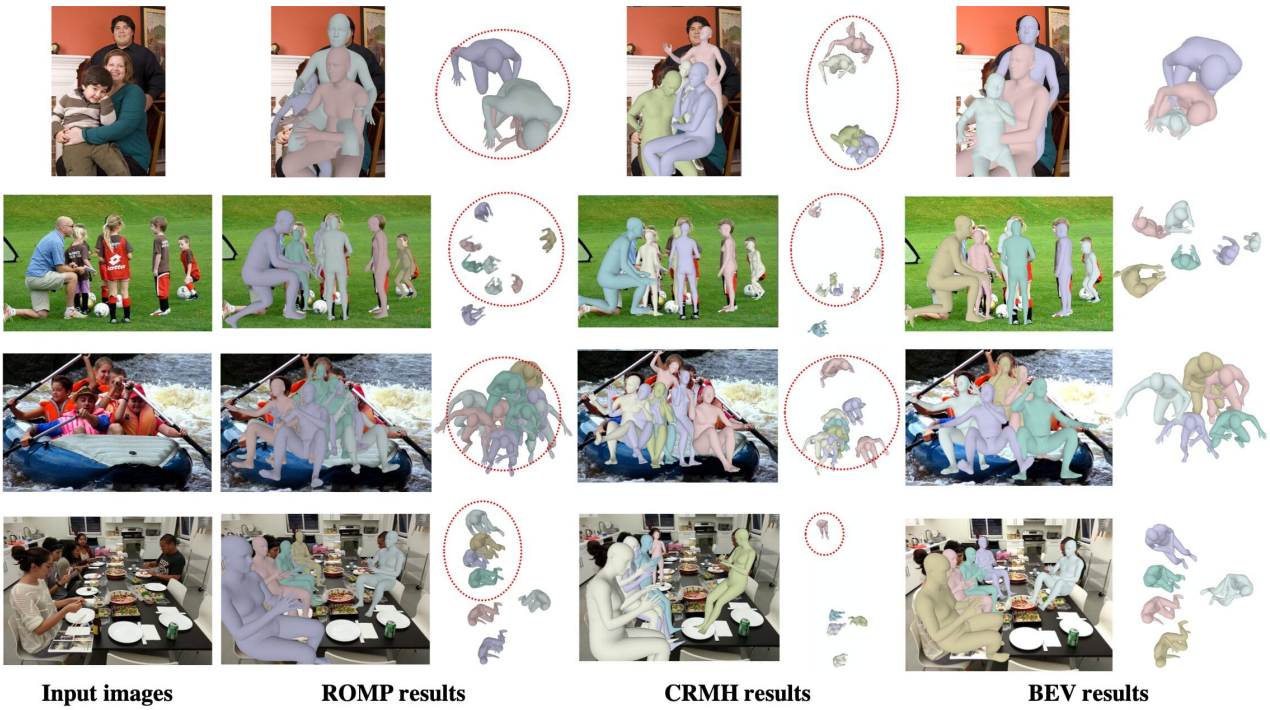


Figure 7. Qualitative comparisons to SOTA methods, ROMP [30] and CRMH [9], on RH test set.

reconstruction. It demonstrates that learning weak annotations via the proposed weakly-supervised training (WST) helps to alleviate the age bias. Multi-stage methods, like Pose2Pose [26], benefit from taking high-resolution person crops as input, which helps process the small-scale subjects in AGORA. Besides, as a sanity check, we also compare with SOTAs on 3DPW and MuPoTS datasets. While not tuned for uncrowded scenes, BEV is on par with the previous methods on MuPoTs (Tab. 4) and 3DPW (Tab. 5).

Method	All↑	Matched↑
CRMH [9]	69.1	72.2
ROMP [30]	69.9	74.6
3DCrowdNet [4]	**72.7**	73.3
BEV	70.2	**75.2**

Table 4. Comparisons to the SOTAs on MuPoTS.

Method	PMPJPE	MPJPE	MPVE
HybrIK [19]	48.8	80.0	94.5
METRO [20]	47.9	77.1	**88.2**
ROMP [30]	47.3	**76.7**	93.4
BEV	**46.9**	78.5	92.3

Table 5. Comparisons to the SOTAs on 3DPW test set.

Method	F1 score↑	MVE↓	MPJPE↓	NMVE↓	NMJE↓
ROMP [30]	0.695	173.76	170.55	249.96	245.34
BEV	**0.732**	**169.21**	**165.27**	**231.16**	**225.76**
w/o WST	0.738	171.16	168.12	235.06	230.89
w/o DC	0.741	170.59	168.12	229.98	225.67

Table 6. 3D mesh/pose error on AGORA-PC, the high occlusion (over 70%) subset of the AGORA validation set (no kids).

5.2. Ablation Studies

To analyse the performance gain of different designs, we perform more ablation studies on AGORA−PC, a high occlusion (over 70%) subset of the AGORA validation set (no kids). This has ground truth 3D annotations for detailed evaluation while the test set does not. BEV uses the same training samples as [30]. Comparing BEV and BEV w/o WST in Tab. 6 also shows that our gains in high occlusion situations come from the 3D representation.

Besides, we also evaluate the effectiveness of depth encoding (DC) for 3D mesh parameter regression. Depth encoding is developed to transfer people at different depths to individual feature spaces. Tab. 6 shows that adding the depth encoding reduces mesh reconstruction error under high occlusion (over 70%). It demonstrates that achieving depth-aware mesh regression via adding depth encoding is beneficial to alleviating depth ambiguity and improving the stability under occlusion.

5.3. Qualitative Results

In Fig. 6, we present more qualitative results on CMU Panoptic and AGORA. In Fig. 1, 7, 8, we show the results under various crowded scenarios, including queuing, standing side by side, and mixed scenarios. Compared with ROMP [30] and CRMH [9], BEV performs much better in detection, depth reasoning, and robustness to occlusion, especially in cases containing children. These results demonstrate the superiority of our 3D representation, WST, and perspective camera model. However, we also observe some limitations of BEV from failure cases in Fig. 9. Without modeling the contact between multiple people, BEV may miss obvious contact and cannot avoid mesh intersections. Besides, BEV is unable to handle occlusions with few visible parts and dense small-scale subjects in crowds.

References

[1] Pexels. https://www.pexels.com. 1, 4, 9, 10

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 3, 4

[3] 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*, pages 5386–5395, 2020. 2

[4] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *CVPR*, 2022. 6

[5] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, pages 20–40, 2020. 5

[6] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *3DV*, pages 792–804, 2022. 5

[7] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[8] 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*, 36(7):1325–1339, 2013. 3

[9] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 4, 6, 7, 9

[10] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, pages 1465–1472, 2011. 3

[11] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342, 2015. 5, 6

[12] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *ECCV*, 2020. 3, 5

[13] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 5

[14] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 5

[15] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *ICCV*, pages 11035–11045, 2021. 4, 5

[16] 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*, pages 2252–2261, 2019. 4, 5

[17] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, pages 11977–11986, 2019. 4

[18] 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*, pages 10863–10872, 2019. 3, 4

[19] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, 2021. 7

[20] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 7

[21] 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*, pages 740–755, 2014. 3, 4

[22] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, pages 5137–5146, 2018. 5

[23] 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*, pages 120–130, 2018. 3, 4

[24] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *TOG*, pages 1–14, 2017. 4

[25] 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 *CVPR*, pages 10133–10142, 2019. 3

[26] Gyeongsik Moon and Kyoung Mu Lee. Pose2Pose: 3d positional pose-guided 3d rotational pose prediction for expressive 3d human pose and mesh estimation. *arXiv*, 2020. 5, 6

[27] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, 2021. 2, 3, 4, 5, 6

[28] 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*, pages 10975–10985, 2019. 5

[29] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV*, pages 1749–1759, 2021. 5

[30] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 1, 2, 4, 5, 6, 7, 9

[31] 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*, pages 601–617, 2018. 4

[32] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, pages 11446–11456, 2021. 5

[33] 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*, pages 889–898, 2019. 4

Figure 8. Qualitative comparisons to SOTA methods, ROMP [30] and CRMH [9], on Internet images [1].

Figure 9. Failure cases on Internet images [1]