Zolly: Zoom Focal Length Correctly for Perspective-Distorted Human Mesh Reconstruction Supplementary Material

Wenjia Wang[1,4] Yongtao Ge[2] Haiyi Mei[3] Zhongang Cai[3,4]
Qingping Sun[3] Yanjun Wang[3] Chunhua Shen[5] Lei Yang[3,4,†] Taku Komura[1]

[1] The University of Hong Kong [2] The University of Adelaide [3] SenseTime Research
[4] Shanghai AI Laboratory [5] Zhejiang University

A. Details of Perspective-distorted Datasets.

A.1. PDHuman

Our pipeline is inspired by recent works on synthetic data [15, 4, 17]. A photogrammetry-scanned human model with a unique body pose will be rendered with a random viewpoint in an HDRi background. The detailed statistics of PDHuman are illustrated in Tab. 1.

Protocol	Train	5	4	3	2	1
Number	126198	2821	4166	6601	12225	27448
τ	1.0	3.0	2.6	2.2	1.8	1.4
Mean f	245	174	176	180	191	230
Mean FoV	98.6°	113.7°	112.0°	112.0°	110.0°	102.0°
Mean T_z	1.3	0.8	0.8	0.8	0.9	1.1

Table 1: Statistical information of PDHuman. τ denotes maximum diisortion scale in the main text.

Human model. We use a corpus of 630 photogrammetry-scanned human models from Renderpeople [3], with well-fitted SMPL parameters. Initially, the body pose is sampled from a collection of high-quality motion sequences obtained from Mixamo [1]. The the pose is converted to SMPL skeleton using a re-targeting approach. Finally, we use a SGD optimizer to optimize the chamfer distance between the SMPL vertices and RenderPeople vertices to refine the pose and shape parameter.

Camera. In order to simulate a wide range of real-world scenarios, a perspective camera is randomly sampled with a focal length that spans from 7mm to 102mm. The corresponding FoV angle is from 10°to 140°. The human mesh is then positioned at the center of a sphere, whose radius is chosen randomly and dependent on the camera's focal length. The camera, facing the center of the sphere, is then placed on the surface of the sphere with a randomized elevation and azimuth angle.

*LY† is the corresponding author.

Figure 1: More examples from PDHuman dataset.

Rendering pipeline. To increase the diversity of data, each frame contains ambient lighting calculated by path tracing in Blender [5] and diverse background generated by HDRi images from PolyHaven [2]. The size of all rendered images is 512×512.

A.2. SPEC-MTP

SPEC-MTP [11] is a real-world image dataset with calibrated focal lengths and well-fitted SMPL parameters (including θ, β), and translation. The images were taken at relatively close-up distances, leading to noticeable perspective distortion in the limbs and torsos of subjects. We use it as one of the evaluation datasets for our task. The detailed

Protocol	3	2	1
Number	713	2609	6083
τ	1.8	1.4	1.0
Mean f	935	976	1114
Mean FoV	69.3°	67.7°	62.4°
Mean T_z	1.1	1.1	1.4

Table 2: Statistical information of SPEC-MTP [11]. τ denotes the maximum distortion scale in the main text.

Protocol	Train	3	2	1
Number	84170	926	2696	6550
τ	1.0	1.8	1.4	1.0
Mean f	318	318	318	318
Mean FoV	127	127	127	127
Mean T_z	1.9	1.9	1.9	1.9

Table 3: Statistical information of HuMMan [6]. τ denotes the maximum distortion scale in the main text.

statistics of SPEC-MTP are illustrated in Tab. 2.

A.3. HuMMan

The HuMMan dataset, as proposed in [6], is a real-world image dataset that utilizes 10 calibrated RGBD kinematic cameras to capture shots for each frame. From these shots, segmented point clouds are extracted from depth images, resulting in a comprehensive dataset. The SMPL parameters were registered upon triangulated 3D keypoints and point clouds, providing ground-truth data that is highly useful for mocap tasks. We reshape all the images to 360×640 pixels. The detailed statistics of HuMMan are illustrated in Tab. 3.

B. Analysis of 3DPW dataset

We divide the 3DPW dataset into three protocols based on the maximum distortion scale τ in Tab. 4. We report the results of our re-implemented HMR-R50 [9] and Zolly-H48 on each protocol in Tab. 5. The experiments indicate that Zolly outperforms HMR-R50, and this improvement is more pronounced as the distortion scale increases. This observation serves as compelling evidence that Zolly's success on 3DPW can be primarily attributed to its superior performance on distorted images.

Protocol	1	2	3
Number	35115	19016	4657
τ	1.0	1.08	1.16
mean f	1966	1966	1966
mean FoV	52°	52°	52°
mean T_z	4.6	3.5	2.9

Table 4: 3 protocols of 3DPW divided by τ. The larger the value of τ, the greater the degree of distortion.

Protocol/Method	3DPW Test		
	PA-MPJPE↓	MPJPE↓	PVE↓
p1(HMR-R50)	50.2	80.9	94.5
p1(Zolly-H48)	39.8	65.0	76.3
Improvement	+10.4	+15.9	+18.2
p2(HMR-R50)	51.3	82.3	96.2
p2(Zolly-H48)	39.9	64.8	76.8
Improvement	+11.4	+17.5	+19.4
p3(HMR-R50)	58.3	93.9	107.8
p3(Zolly-H48)	44.7	71.7	84.6
Improvement	+ 13.6	+22.2	+23.2

Table 5: Reults of our re-implemented HMR [9] and Zolly-H48 on different protocols of 3DPW. Mainly for showing the correlation between performance improvement and distortion.

C. Details of Model-based Variant of Zolly

As shown in Fig. 2, we introduce a model-based Zolly$^{\mathcal{P}}$ by changing the mesh reconstruction module. Different from Zolly, we regress SMPL parameters rather than 3D vertex coordinates through a transformer decoder in Zolly$^{\mathcal{P}}$. We warp the grid Feature F_{grid} into UV space (F_{grid-w}) via I_{IUV} to eliminate the spatial distortion of each part of the features, then concatenate the warped distortion feature F_{grid-w} and regress SMPL parameters from it. We represent the 24 rotations of joints θ and body shape parameters β as 25 learnable tokens. The translation estimation module and supervision are exactly the same as Zolly.

D. More about Cameras

Affine Transformation. Our affine transformation of translation is the same as SPEC [11]. T_x, T_y and t_x, t_y should satisfy the following equation for every x, y by connecting the re-projected coordinates in the cropped image coordinate system and original image coordinate system in screen space.

$$\begin{bmatrix} \frac{w}{2}\left[s_x(x+t_x))+1\right]+c_x-\frac{w}{2} \\ \frac{h}{2}\left[s_y(y+t_y))+1\right]+c_y-\frac{h}{2} \end{bmatrix} = \begin{bmatrix} \frac{W}{2}\left[S_W(x+T_x)+1\right] \\ \frac{H}{2}\left[S_H(y+T_y)+1\right] \end{bmatrix}. \tag{1}$$

where $S_W = s_x/(\frac{W}{w})$, $S_H = s_y/(\frac{H}{h})$, so we can get the transform by:

$$\begin{bmatrix} T_x \\ T_y \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & (2c_x - W)/ws_x \\ & 1 & (2c_y - H)/hs_y \\ & & 1 \end{bmatrix} \begin{bmatrix} t_x \\ t_y \\ 1 \end{bmatrix} \tag{2}$$

c_x, c_y terms the bounding-box center coordinate in original image. h, w terms the cropped image size, where H, W terms original image size. During training, we will expand every bounding box of the human body to a square and resize the cropped image to 224×224 pixels.

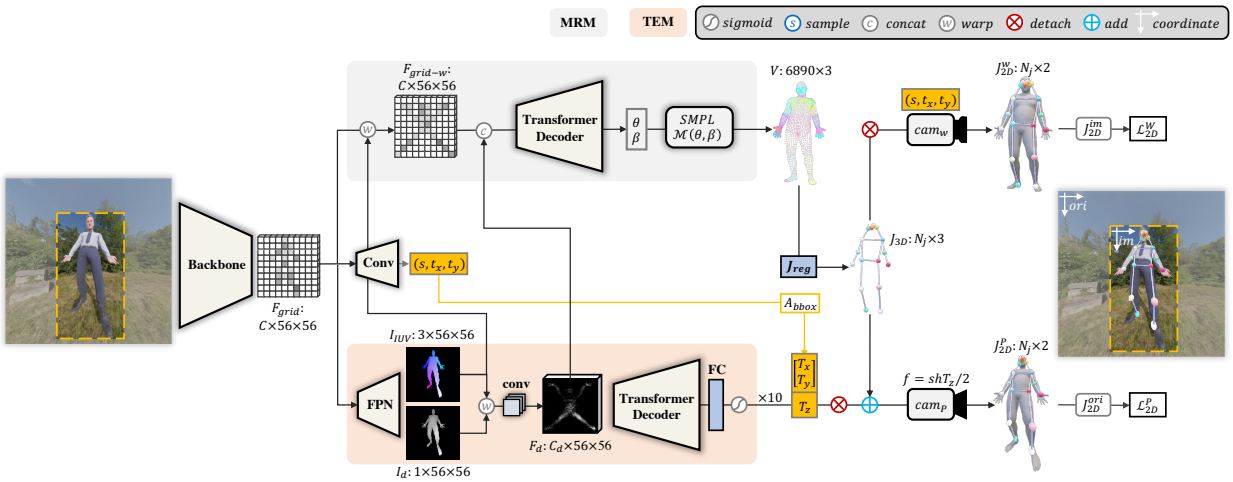

Figure 2: Zolly$^{\mathcal{P}}$ **pipeline overview**. Compared to Zolly, the main difference is the reformulated mesh reconstruction module.

T_z	0.5	0.75	1.0	2.0	4.0	8.0	12.0	16.0	20
τ	3.06	1.69	1.41	1.16	1.07	1.04	1.02	1.02	1.01
$Error$	30.56	10.46	7.38	3.54	1.76	0.88	0.59	0.44	0.35

Table 6: Distortion and re-projection error caused by distance in CMU-MOSH [13]. The re-projected error is measured in pixels.

Analysis of dolly zoom. The dolly zoom is an optical effect performed in-camera, whereby the camera moves towards or away from a subject while simultaneously zooming in the opposite direction. It was first proposed in the film JAW [16]. In this section, we simulate the effect on CMU-MOSH [13] data. First, we get all the vertices in CMU-MOSH [13] by feeding the SMPL [14] parameters to the body model. We further obtain 3D joints by multiplying the joint regressor matrix to the vertices. We then apply weak-perspective camera parameters (s, t_x, t_y) while adjusting the distance to approximate the human body's location and size, producing increased distortion as the camera approaches. The weak-perspectively projected 2D joints are $s(x + t_x, y + t_y)$, where x, y are the corresponding 3D joint coordinates. We set the image height to 224 pixels and re-project the 2D joints and compare the error between the weak-perspective and perspective projection results. As shown in Tab. 6, when the subject is located over 4 meters away, the re-projected error is only 1.76 pixels, which is negligible on a 224-pixel image. When the subject is further than 8 meters, the error is less than 1 pixel, indicating non-distortion of the images.

E. More quantitative results

Full results on PDHuman: As shown in Tab. 7 and Tab. 8, we report results on all 5 protocols in the PDHuman test dataset. Our proposed methods, Zolly (H48) and Zolly$^{\mathcal{P}}$

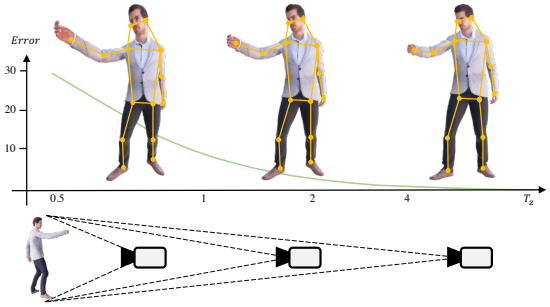

Figure 3: Distortion and re-projection error caused by distance. The vertical axis is measured in pixels, and the horizontal axis is measured in meters.

(R50) outperform the other methods in all metrics by a large margin.

Full results on SPEC-MTP: As illustrated in Tab. 9, we report the results of all the 3 protocols in SPEC-MTP dataset. In this real-world dataset, Zolly (H48) largely outperforms other methods in all metrics. In column $\tau = 1.0$, Note that our re-implemented SPEC$*$ achieves higher performance than the official implementation.

Full results on HuMMan: As shown in Tab. 10, Zolly (H48) largely outperforms other methods in all metrics. By contrast, although CLIFF [12] performs comparably well on the HuMMan dataset, it demonstrates poor performance on the PDHuman dataset. We conjecture the focal length assumption of CLIFF is suitable for datasets captured by fixed and similar camera settings, *e.g.* HuMMan dataset, while not valid for the PDHuman dataset with varied camera settings.

F. Qualitative results.

Qualitative results on Human3.6M [8] dataset. We show qualitative results of Zolly on Human3.6M dataset in Fig. 4

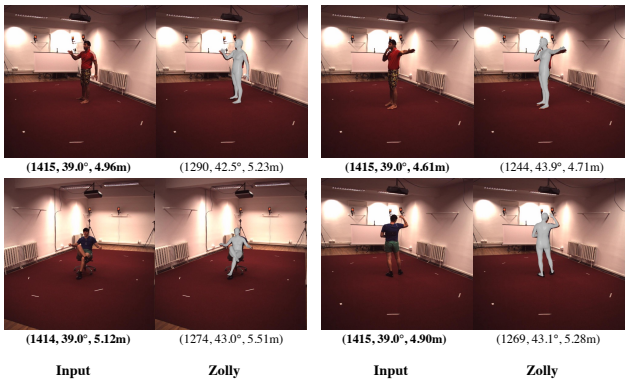

Figure 4: Qualitative results on Human3.6M dataset. The number under each image represents predicted/ground-truth focal length f, FoV angle, and z-axis translation T_z. Our method could predict an approximate translation for non-distorted images as well.

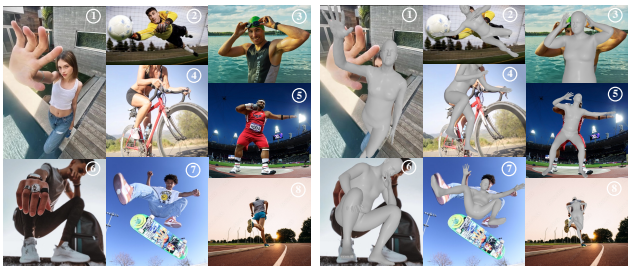

Figure 5: Failure cases. The left part is input, and the right part is our prediction.

Failure cases. Although our methodology is generally effective, it has trouble under certain extreme circumstances. As demonstrated in Fig. 5, due to the lack of training data containing characters with large hands ((1), (2), and (6)), and large feet ((7) and (8)), Zolly produce sub-optimal results on such images. Similarly, our approach may not perform well on characters with exceptional body shapes, as exemplified by (2), (3), and (7), where the athletes have muscular bodies. Additionally, it is difficult for Zolly to reconstruct self-occluded human bodies, as depicted in (4). We are actively exploring strategies to address these limitations and improve the robustness of our methodology.

More qualitative results on distorted images. We show more qualitative results of Zolly comparing with SOTA methods for perspective-distorted images on PDHuman (Fig. 6), Web images (Fig. 7), and SPEC-MTP (Fig. 8).

Methods	PDHuman ($\tau = 3.0$)					PDHuman ($\tau = 2.6$)					PDHuman ($\tau = 2.2$)				
	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑
HMR (R50) [9]	62.5	91.5	106.6	48.9	21.7	59.9	87.8	102.4	50.0	22.5	57.4	84.0	98.1	51.4	23.6
HMR-f (R50) [9]	61.6	90.2	105.5	45.2	20.4	59.2	86.6	101.3	46.5	21.4	56.8	82.9	97.2	48.1	22.7
SPEC (R50) [11]	65.8	94.9	109.6	43.4	19.6	63.2	91.5	105.8	43.3	19.5	60.6	87.3	101.3	42.2	18.7
CLIFF (R50) [12]	66.2	99.2	115.2	51.4	24.8	63.4	94.4	109.8	52.7	25.9	60.6	89.6	104.3	54.2	27.1
PARE (H48) [10]	66.3	95.9	116.7	48.2	20.9	63.6	92.3	112.7	49.3	21.7	60.6	88.7	108.6	50.7	22.7
GraphCMR (R50)	62.1	85.8	98.4	47.9	21.5	59.5	82.6	94.8	49.1	22.4	56.8	78.8	90.4	50.5	23.6
FastMetro(H48) [7]	58.6	83.6	95.4	50.1	22.5	55.8	79.9	91.4	51.4	23.5	53.1	75.9	86.7	52.9	24.9
Zolly^{P} (R50)	54.3	80.9	93.9	**54.5**	**27.4**	52.4	77.5	90.2	**55.7**	28.5	50.0	74.0	86.4	**56.9**	**29.5**
Zolly (R50)	54.3	76.4	87.6	51.4	24.0	51.8	73.3	84.1	52.4	24.8	49.3	70.1	80.6	53.3	25.7
Zolly (H48)	**49.7**	**70.2**	**81.2**	50.5	23.8	**47.6**	**64.3**	**74.4**	55.3	**28.5**	**44.9**	**64.3**	**74.7**	55.3	28.5

Table 7: Results of SOTA methods on PDHuman ($\tau = 3.0$, $\tau = 2.6$, $\tau = 2.2$ protocols). HMR-f terms HMR [9] model trained with same focal length as Zolly.

Methods	PDHuman ($\tau = 1.8$)					PDHuman ($\tau = 1.4$)				
	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑
HMR (R50) [9]	53.9	79.0	92.4	53.6	25.1	49.2	73.3	85.9	57.3	28.2
HMR-f (R50) [9]	53.4	78.3	91.8	50.4	24.6	48.8	72.7	85.3	54.6	28.1
SPEC (R50) [11]	56.8	81.8	95.1	40.1	17.1	51.8	75.4	87.9	37.4	15.3
CLIFF (R50) [12]	56.7	83.6	97.3	56.5	29.1	51.6	76.9	89.7	60.2	32.7
PARE (H48) [10]	56.8	83.9	103.0	52.8	24.3	51.8	78.5	96.6	56.6	27.5
GraphCMR (R50)	53.2	74.2	85.2	52.7	25.3	48.7	69.1	79.4	56.4	28.6
FastMetro(H48) [7]	49.4	71.1	81.1	55.5	27.0	45.0	65.8	75.2	59.7	31.0
Zolly^{P} (R50)	47.1	69.8	81.6	**58.7**	**30.7**	43.2	65.2	76.5	**61.3**	**32.6**
Zolly (R50)	45.9	66.0	75.9	54.8	26.8	41.9	61.5	70.9	57.2	28.2
Zolly (H48)	**42.1**	**60.7**	**70.4**	56.8	29.5	**39.4**	**56.6**	**69.6**	58.3	29.9

Table 8: Results of SOTA methods on PDHuman ($\tau = 1.8$, $\tau = 1.4$ protocols).

Methods	SPEC-MTP ($\tau = 1.8$)					SPEC-MTP ($\tau = 1.4$)					SPEC-MTP ($\tau = 1.0$)				
	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑
HMR (R50) [9]	73.9	121.4	145.6	48.8	16.0	73.1	112.5	135.7	51.1	20.0	69.6	111.8	135.7	50.5	21.8
HMR-f (R50) [9]	72.7	123.2	145.1	52.3	21.0	72.1	113.3	135.5	51.9	21.9	69.1	112.8	136.3	52.5	24.8
SPEC (R50) [11]	76.0	125.5	144.6	49.9	18.8	72.4	114.0	134.3	49.3	19.5	67.4	110.6	132.5	49.1	21.2
SPEC * (R50) [11]	-	-	-	-	-	-	-	-	-	-	71.8	116.1	136.4	-	-
CLIFF (R50) [12]	74.3	115.0	132.4	53.6	23.7	70.2	107.0	126.8	52.0	22.1	67.4	108.7	130.4	51.9	23.4
PARE (H48) [10]	74.2	121.6	143.6	55.8	23.2	71.6	112.7	137.2	55.1	22.4	68.5	113.5	139.6	55.3	25.1
GraphCMR (R50)	76.1	121.1	133.1	56.3	23.4	74.4	114.9	129.5	52.6	20.8	70.2	112.7	127.8	51.7	22.0
FastMetro(H48) [7]	75.0	123.1	137.0	53.5	20.5	70.8	112.3	128.0	52.4	20.6	66.3	110.2	126.5	51.8	22.6
Zolly^{P} (R50)	72.9	117.7	138.2	54.7	22.4	70.5	108.1	129.4	53.9	21.5	68.4	110.2	134.3	54.7	24.2
Zolly (R50)	74.0	122.1	135.6	58.9	24.9	70.3	111.1	126.0	56.9	22.0	66.9	109.6	124.4	56.5	23.4
Zolly (H48)	**67.4**	**114.6**	**126.7**	**62.6**	**30.4**	**66.5**	**106.1**	**120.1**	**59.9**	**26.6**	**65.8**	**108.2**	**121.9**	**58.5**	**27.0**

Table 9: Results of SOTA methods on SPEC-MTP ($\tau = 1.8$, $\tau = 1.4$, $\tau = 1.0$ protocols). SPEC-MTP ($\tau = 1.0$) indicates the original SPEC-MTP [11] dataset. SPEC * terms the results reported in SPEC [11].

Methods	HuMMan ($\tau = 1.8$)					HuMMan ($\tau = 1.4$)					HuMMan ($\tau = 1.0$)				
	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑
HMR (R50) [9]	30.2	43.6	52.6	65.1	39.5	31.9	45.0	39.5	66.6	39.9	30.0	44.1	50.7	66.6	39.5
HMR-f (R50) [9]	29.9	43.6	53.4	62.7	34.9	31.3	45.0	53.3	66.6	39.9	29.8	44.1	50.7	66.6	39.5
SPEC (R50) [11]	31.4	44.0	54.2	51.4	24.6	33.1	46.1	41.7	46.0	19.2	31.2	44.8	51.6	42.2	16.6
CLIFF (R50) [12]	28.6	42.4	50.2	68.9	44.7	30.3	43.3	51.2	70.2	44.9	28.3	42.3	48.5	70.6	44.5
PARE (H48) [10]	32.6	53.2	65.5	64.5	38.3	33.6	53.3	66.2	65.1	38.0	32.2	53.1	64.6	65.0	37.6
GraphCMR (R50)	29.5	40.6	48.4	61.6	37.5	30.3	40.6	48.2	62.6	37.6	29.3	40.2	46.3	62.8	37.0
FastMetro(H48) [7]	26.3	38.8	45.6	68.3	45.2	27.8	39.9	46.6	69.9	45.7	26.5	38.5	43.6	70.0	45.3
Zolly^{P} (R50)	24.4	36.7	45.9	70.4	**45.5**	26.2	37.6	45.6	70.4	45.3	25.6	37.7	43.7	70.8	45.2
Zolly (R50)	25.5	36.7	43.4	67.0	38.4	25.6	36.5	42.5	70.4	42.7	24.2	35.2	40.4	70.7	42.4
Zolly (H48)	**22.3**	**32.6**	**40.0**	**71.2**	45.1	**24.1**	**33.8**	**40.7**	**72.2**	**47.9**	**23.0**	**33.0**	**38.7**	**73.2**	**47.4**

Table 10: Results of SOTA methods on HuMMan ($\tau = 1.8$, $\tau = 1.4$, $\tau = 1.0$ protocols).

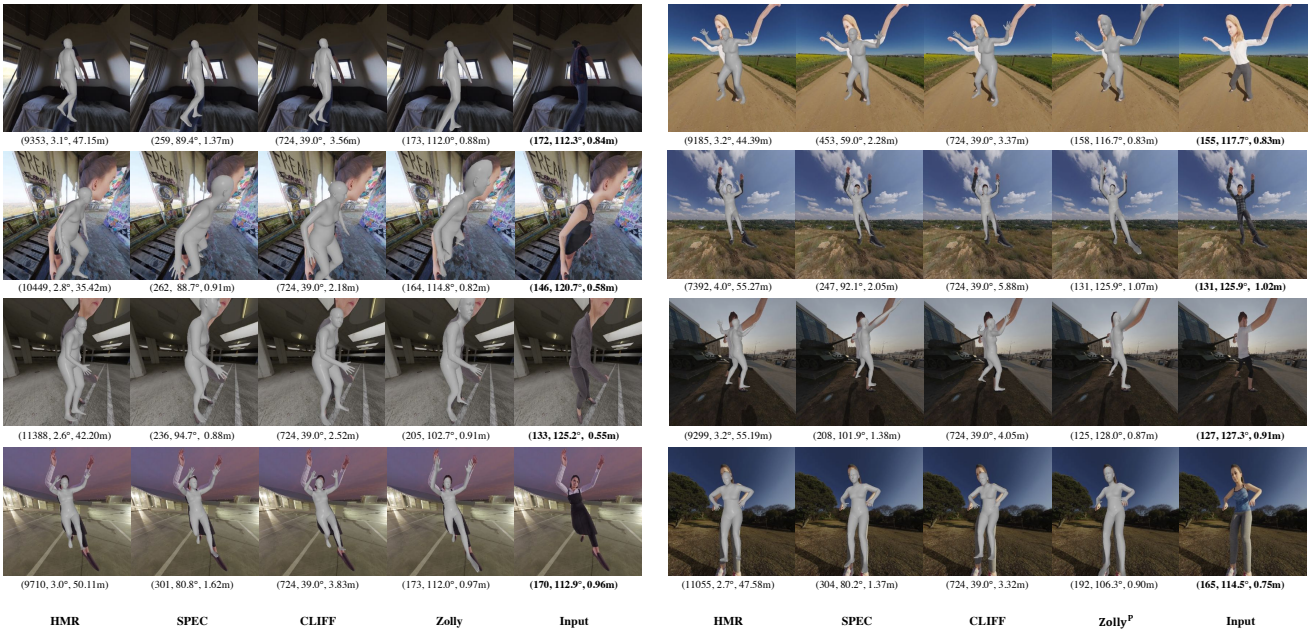

Figure 6: Qualitative results on PDHuman dataset. The number under each image represents predicted/ground-truth focal length f, FoV angle, and z-axis translation T_z.

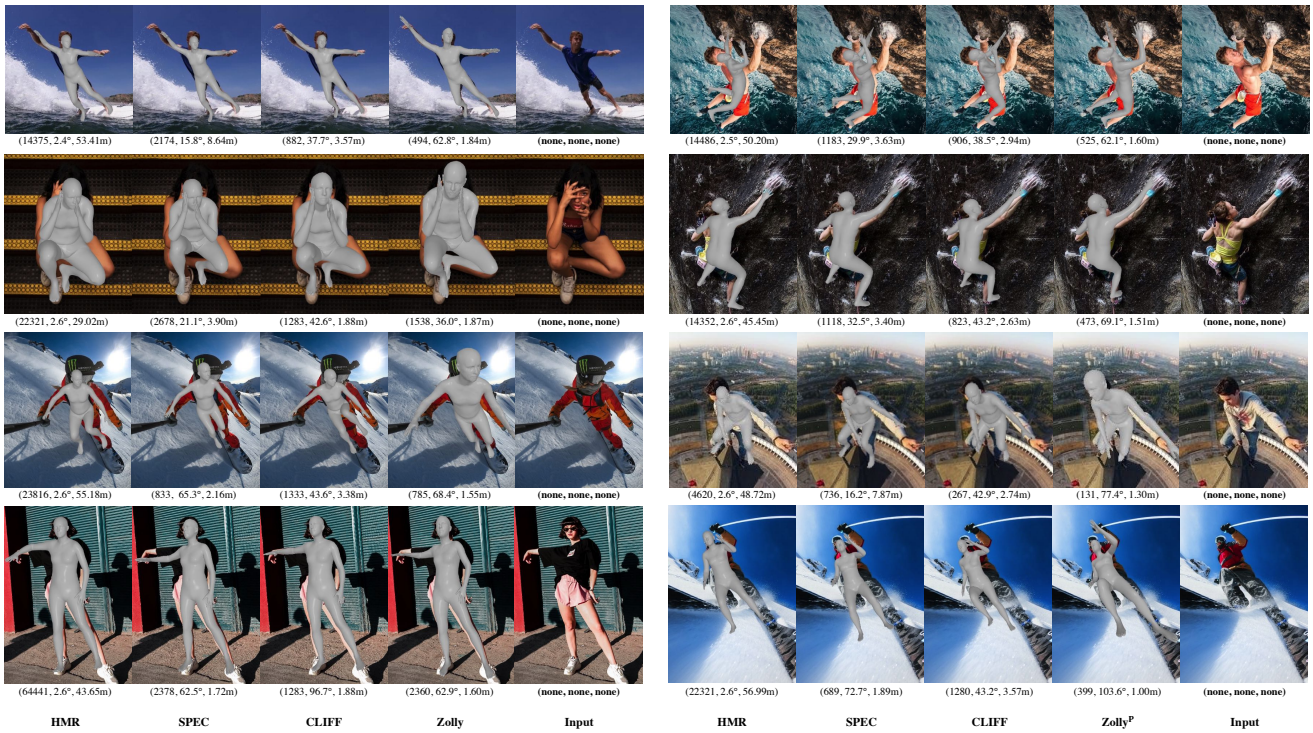

Figure 7: Qualitative results on in-the-wild images. The number under each image represents the predicted focal length f, FoV angle, and z-axis translation T_z. Images are collected from https://pexels.com and https://yandex.com.

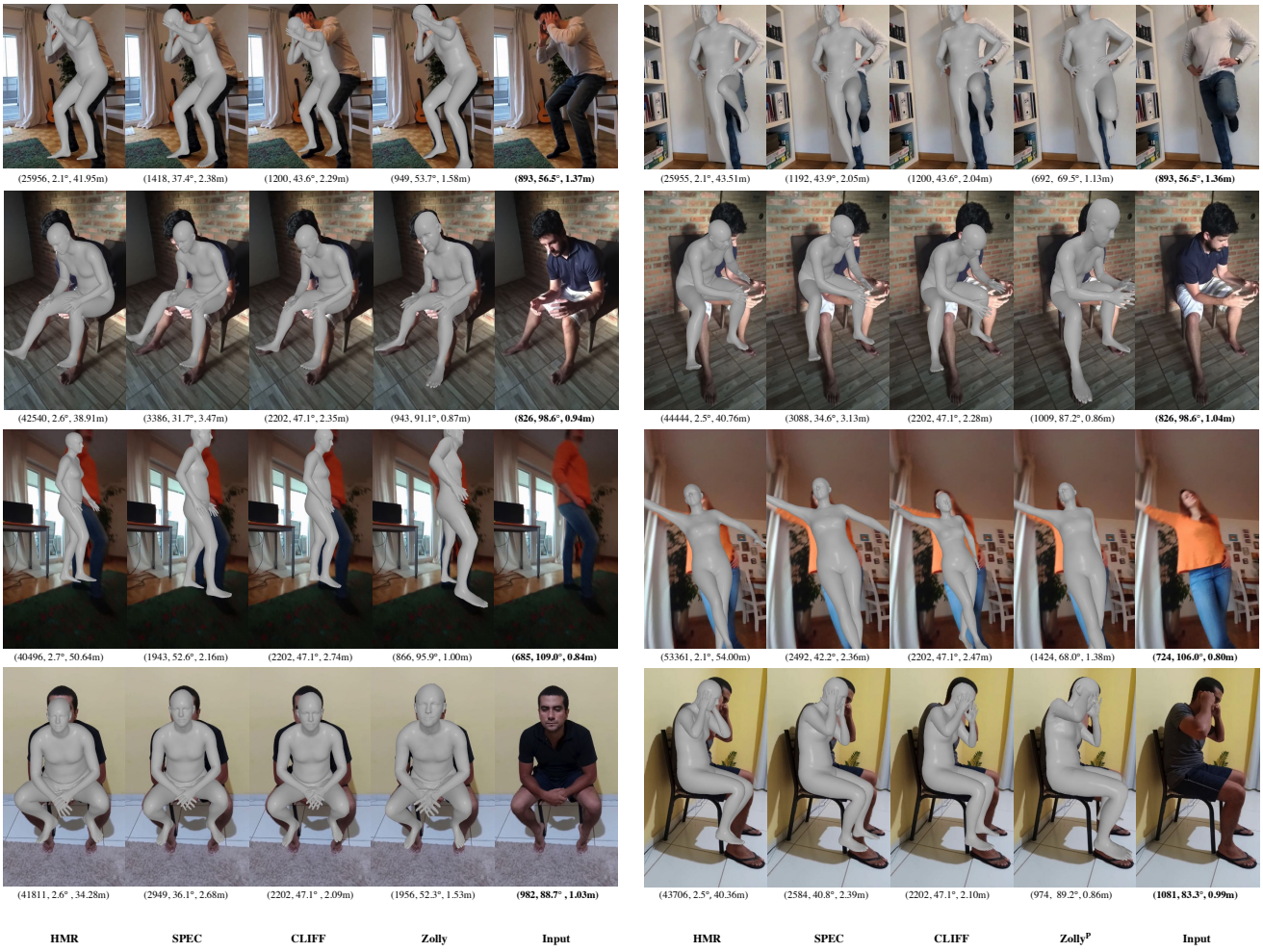

Figure 8: Qualitative results on SPEC-MTP dataset. The number under each image represents predicted/ground-truth focal length f, FoV angle, and z-axis translation T_z. The ground-truth T_z and focal length f for SPEC-MTP are pseudo labels.

References

[1] *Mixamo*, 2022. https://www.mixamo.com. 1

[2] *PolyHaven*, 2022. https://polyhaven.com. 1

[3] *Renderpeople*, 2022. https://renderpeople.com. 1

[4] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Hspace: Synthetic parametric humans animated in complex environments. *arXiv: Comp. Res. Repository*, 2021. 1

[5] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2022. 1

[6] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *Eur. Conf. Comput. Vis.*, pages 557–577. Springer, 2022. 2

[7] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *Eur. Conf. Comput. Vis.*, pages 342–359. Springer, 2022. 5

[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. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 4

[9] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7122–7131, 2018. 2, 5

[10] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Int. Conf. Comput. Vis.*, pages 11127–11137, 2021. 5

[11] 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 *Int. Conf. Comput. Vis.*, pages 11035–11045, 2021. 1, 2, 5

[12] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. *arXiv: Comp. Res. Repository*, 2022. 3, 5

[13] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Mosh: Motion and shape capture from sparse markers. In *ACM Transactions on Graphics (TOG)*, volume 33, pages 220:1–220:13. ACM, 2014. 3

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):1–16, 2015. 3

[15] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2021. 1

[16] Steven Spielberg. Jaws, June 1975. 3

[17] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1