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

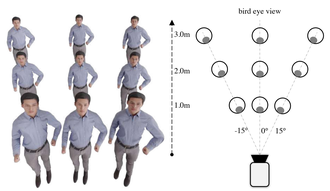
Wenjia Wang [1,2] Yongtao Ge [3] Haiyi Mei [4] Zhongang Cai [4]
Qingping Sun [4] Yanjun Wang [4] Chunhua Shen [5] Lei Yang [2,4] Taku Komura [1]

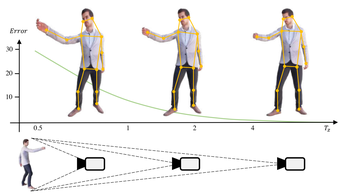
[1]The University of Hong Kong [2]Shanghai AI Laboratory [3]The University of Adelaide [4]SenseTime Research [5]The Zhejiang University

ICCV23 PARIS

Introduction

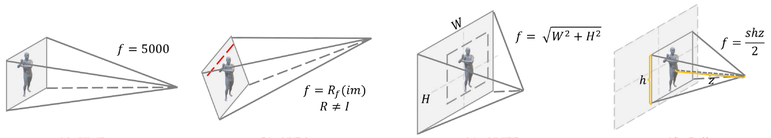
- **Motivation:** Existing 3D human mesh reconstruction methods use a constant f or estimate one based on the background context. Such f deviates a lot from distorted images caused by perspective projection.
  The distortion is directly caused by the distance and the facing angle to the camera center.
  Close-view shots could cause distortion on human bodies, which could be used to calculate the f.





- **Contribution**
  A novel camera system for the perspective-distorted 3DHMR task.
  A new neural model, and a hybrid re-projection loss.
  A new and the first synthetic dataset, PDHuman, for perspective-distorted 3D human pose estimation.

Camera System Design



(a). HMR (b). SPEC (c). CLIFF (d). Zolly

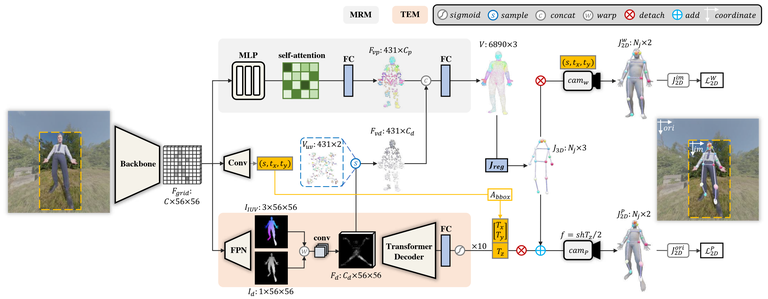
- **HMR**. $f = 5000$. Most methods follow this setting.
- **SPEC**. The f is estimated by a network pre-trained on other datasets.
- **CLIFF**. Use the diagonal length as f if no ground truth f.
- **Zolly**. $f = shT_z/2$. Where T_z is the z-axis distance.

The weak-perspective camera parameters (s, t_x, t_y), which represent 2D orthographic transformation, could be used to approximate the projection:

$$\begin{bmatrix} f(x+T_x)/T_z \\ f(y+T_y)/T_z \end{bmatrix} = \begin{bmatrix} s(x+t_x) \\ s(y+t_y) \end{bmatrix}, s \times T_z = f, T_x = t_x, T_y = t_y. \quad (1)$$

In previous methods, they either use a constant or estimated f, and calculate distance by $T_z = f/s$. On contrary, we estimate T_z based on human body distortion clues and calculate f by $f = s \times T_z$. (NDC space)

Proposed Pipeline



- **Translation Estimation Module.** We use a Transformer to regress the z-axis distance from the warped I_{IUV} image and use sigmoid then ×10 to restrict T_z less than 10m.
- **Mesh Estimation Module.** We adopt an MLP structure to predict the coordinates of a coarse mesh of the body, then up-sample the mesh using two fully connected layers.
- **Loss Functions.** The total loss function is the summation of mesh loss, translation loss, and hybrid re-projection loss.

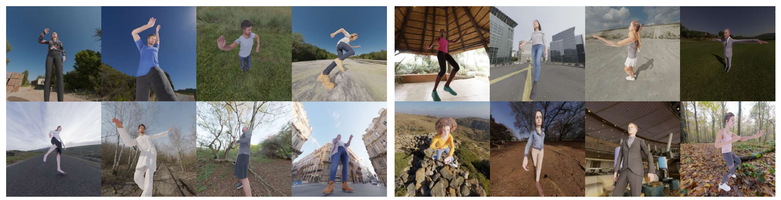
$$\mathcal{L}_{total} = \mathcal{L}_{Mesh} + \mathcal{L}_{Transl} + \mathcal{L}_{2D}^{W} + \mathcal{L}_{2D}^{P} \quad (2)$$

where the $\mathcal{L}_{2D^W}$ is the weak-perspective and $\mathcal{L}_{2D^P}$ is the perspective projection.

New Virtual Dataset: PDHuman

We propose the first dataset which aims for perspective-distorted 3D human pose estimation.

- **Amount**: 126, 198 images in training and 27, 448 images in testing split.
- **Annotations**: Camera intrinsic matrix, 2D/3D keypoints, SMPL parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and translation.
- **Camera**: Use the dolly-zoom effect to generate random camera intrinsic matrices.
- **Rendering**: Use human models from RenderPeople and body pose sequences from Mixamo, with HDRi images as backgrounds. Use Blender to render the RGB images.



Experiments

- Results on ordinary datasets and distorted datasets.

Method	Backbone	3DPW			Human3.6M	
		PA-JPE	MPJPE	PVE	PA-JPE	MPJPE
HMR	Res50	72.6	116.5	-	56.8	88.0
SPEC	Res50	52.7	96.4	-		
CLIFF	HR48	43.0	69.0	81.2	32.7	47.1
Zolly	HR48	**39.8**	**65.0**	**76.3**	**32.3**	**49.4**

Method	Backbone	PDHuman (p5)			SPEC-MTP (p3)		
		PA-JPE	PVE	mIoU	PA-JPE	PVE	mIoU
HMR	Res50	62.5	106.7	21.7	73.9	145.6	16.0
SPEC	Res50	65.8	109.6	19.6	76.0	144.6	18.8
CLIFF	HR48	66.2	115.2	24.8	74.3	132.4	23.7
Zolly	HR48	**49.9**	**82.0**	**26.5**	**67.4**	**126.7**	**30.4**

Quality Results

Qualitative results of SOTA methods. Row 1: PDHuman test. Row 2, 3, 4: web images. Row 4: SPEC-MTP. Row 6: 3DPW. The number under each image represents predicted/ground-truth f, FoV angle, and T_z. The focal lengths here are all transformed to pixels in full image.