ICCV23
PARIS

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

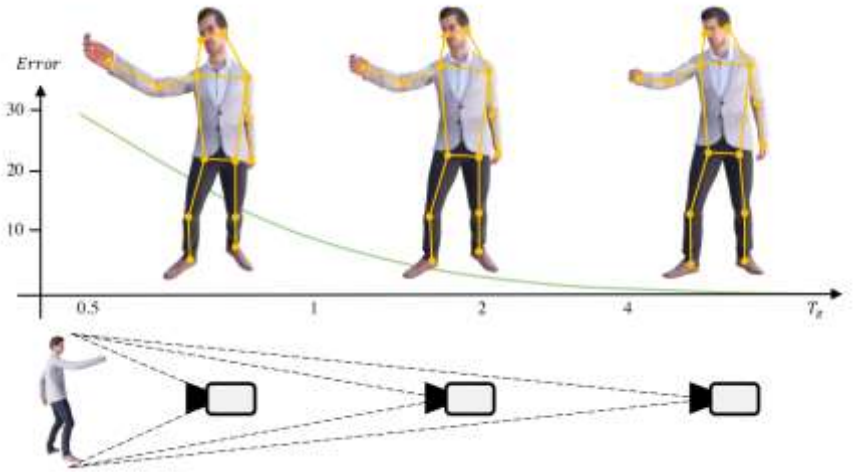
Wenjia Wang[1,2], Yongtao Ge[3], Haiyi Mei[4], Zhongang Cai[4],
Qingping Sun[4], Yanjun Wang[4], Chunhua Shen[5], Lei Yang[2,4], Taku Komura[1]

[1]The University of Hong, [2]Shanghai AI Laboratory

[3]The University of Adelaide, [4]SenseTime Research, [5]The Zhejiang University

Our Findings

- Existing 3D human mesh reconstruction methods either use a constant focal length or estimate one based on the background context.
- Close-view shots could cause distortion on human bodies, which could be used to calculate the distance. When human body is far enough, the large focal length will not hurt the projection.
- This is called dolly zoom. We find the inversion of this phenomenon and use it to solve the relationship between distance and focal length.

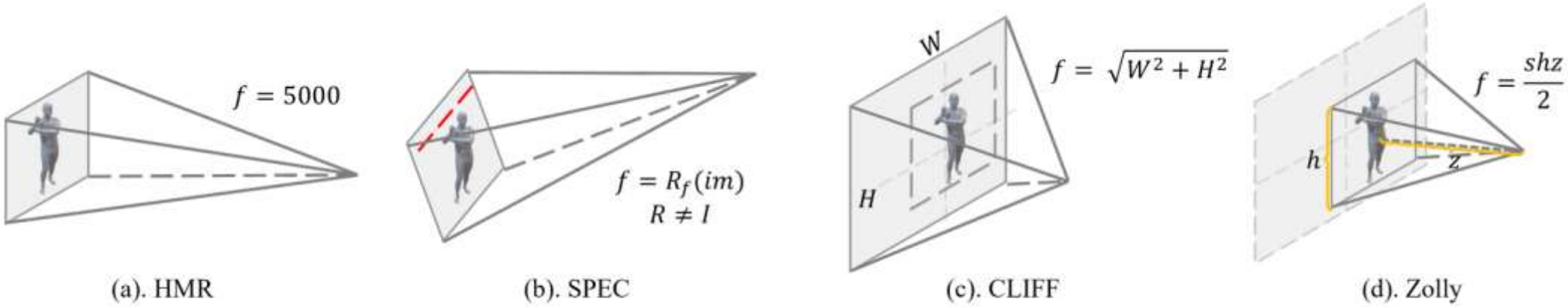


Jaws, a 1975 American thriller film directed by Steven Spielberg

Contribution

- We propose a novel camera system for the perspective-distorted human mesh reconstruction task.
- We introduce a new neural model, consists of a translation estimation module, a mesh reconstruction module, and a hybrid re-projection loss.
- We render a new and the first synthetic dataset, PDHuman, for perspective-distorted 3D human pose estimation.

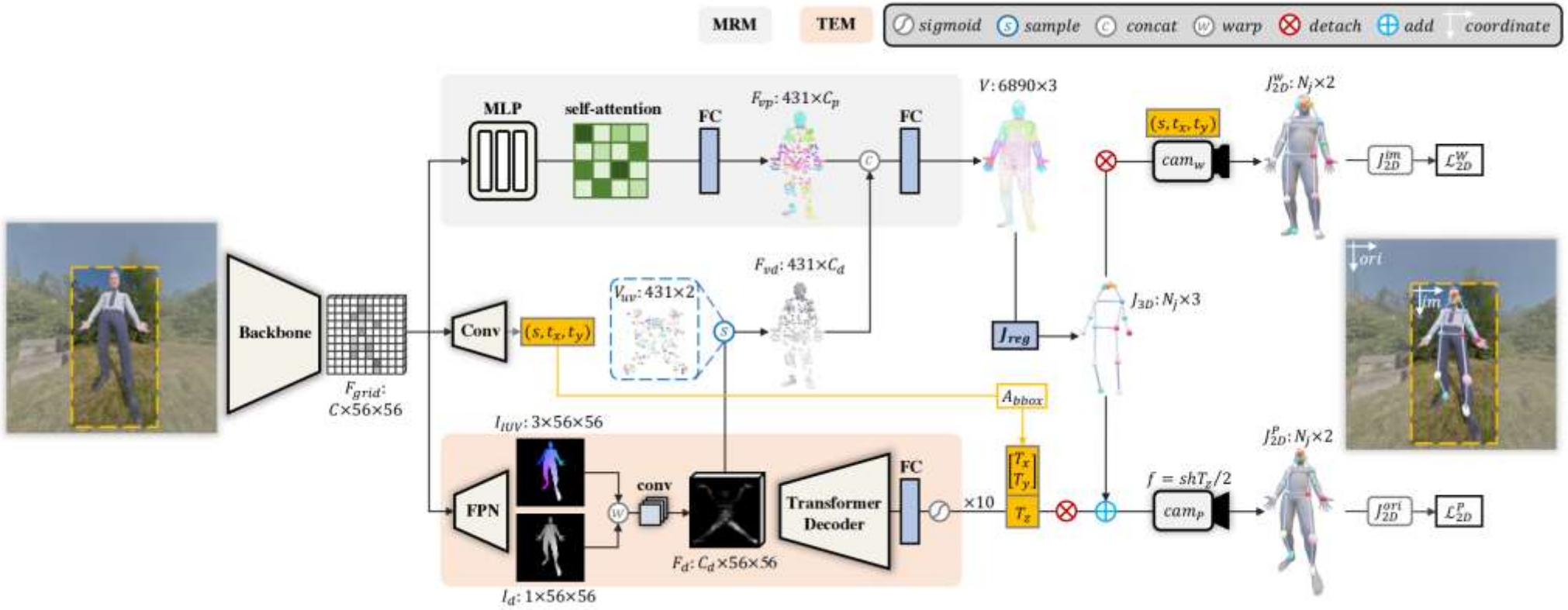
Camera System Design



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

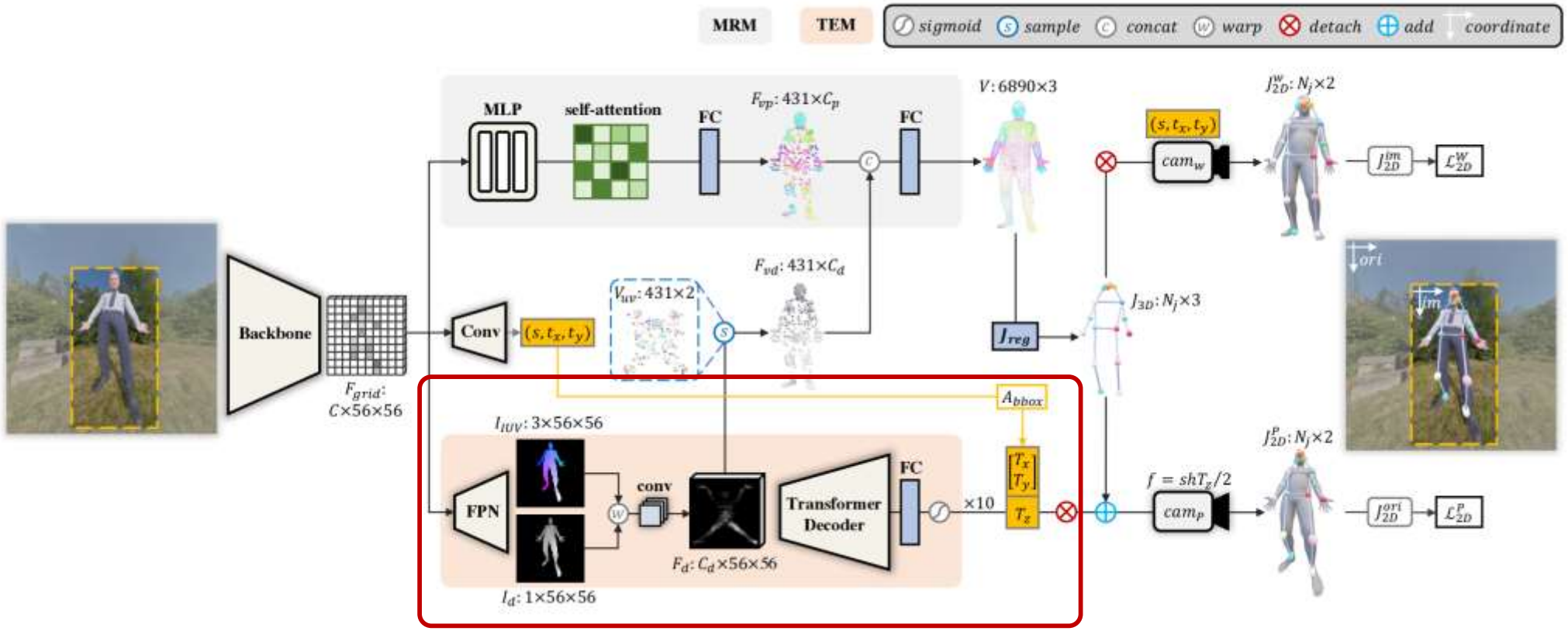
- HMR. $f = 5000\ (pixels)$. Most methods follow this setting.
- SPEC. The f is estimated by a network pre-trained on other datasets.
- CLIFF. Use the diagonal length as f if no ground truth f.
- Zolly. $f = sT_z$.(NDC Space) Where T_z is the z-axis distance. (Please find detailed information in our Paper)

Proposed Method



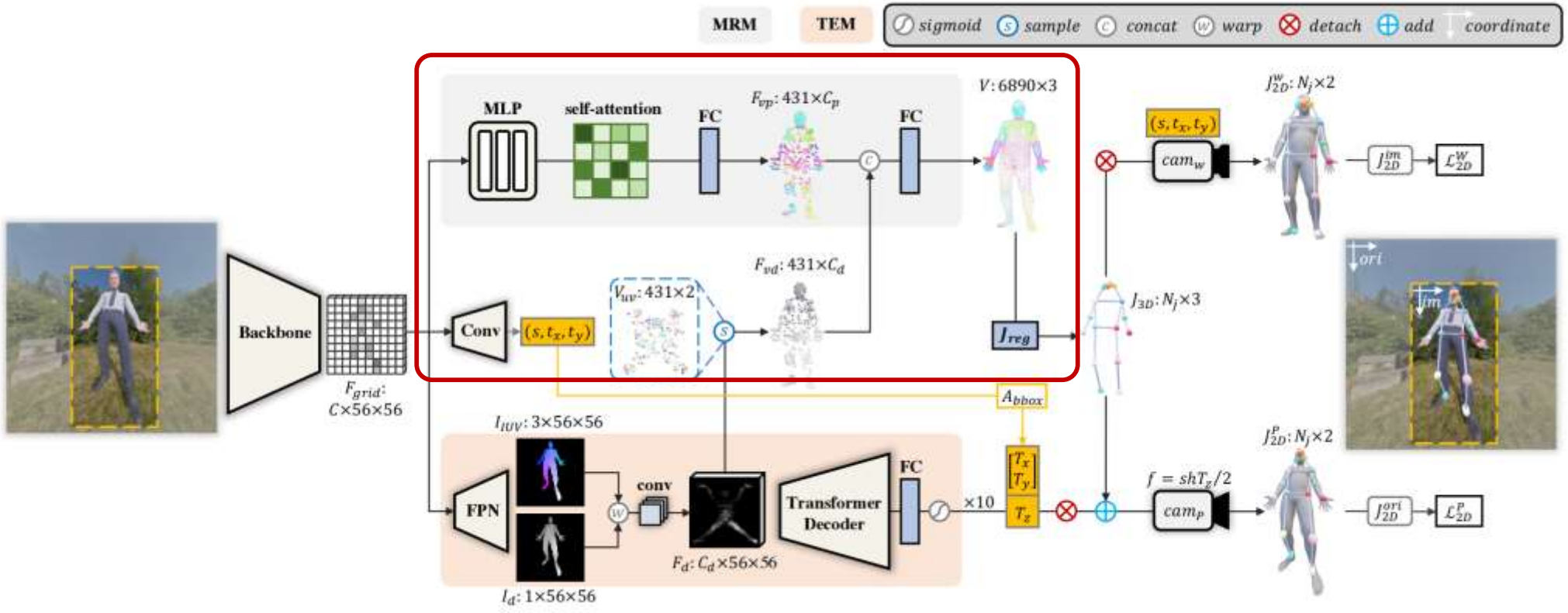
- According to our setting $f = sT_z$, we need to estimate the distance T_z and orthographics scale s. Our pipeline consists of 3 main parts: Translation Estimation Module, Mesh Reconstruction Module, and a hybrid reprojection supervision loss.

Proposed Method



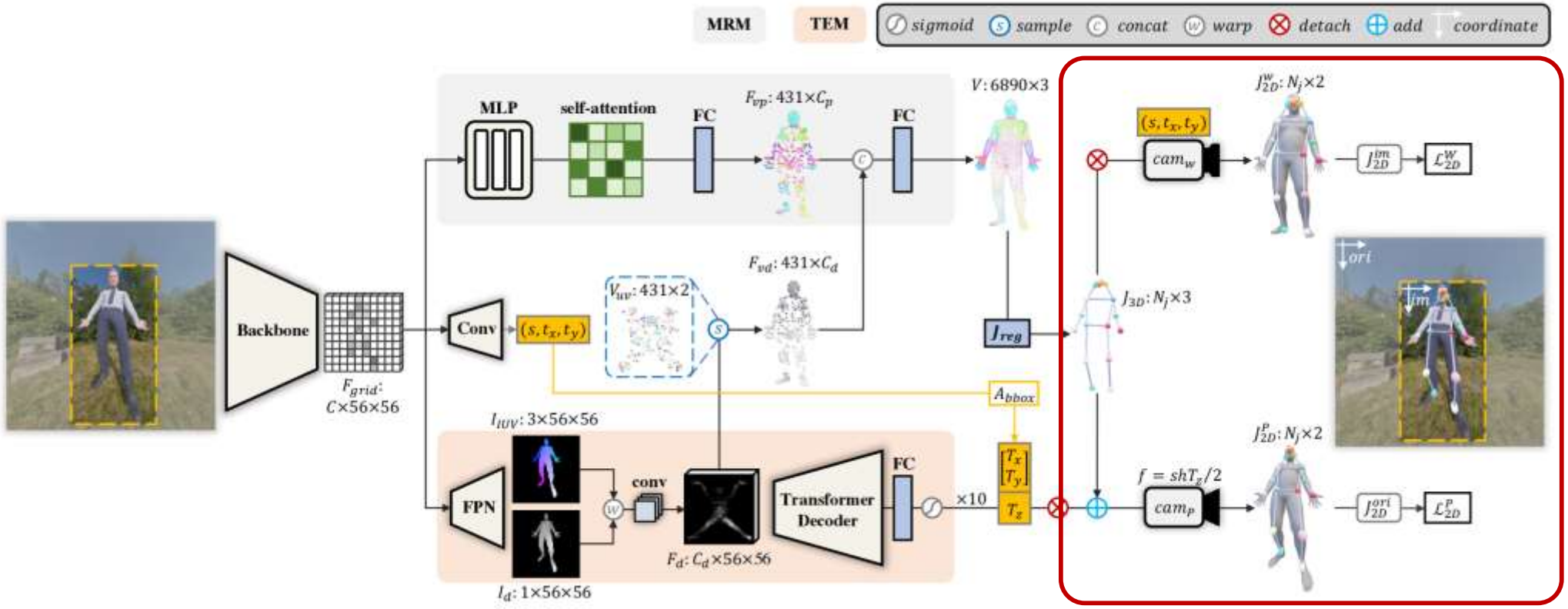
- Translation Estimation Module. We use a Transformer to regress the z-axis distance from the warped IUV image and use $10 \times sigmoid()$ to restrict T_z to be in the range of [0,10m].

Proposed Method



- 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.

Proposed Method



- Loss Functions. The total loss function is the summation of mesh loss, translation loss, and hybrid re-projection loss.

PDHuman Dataset

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

- Annotations: Camera intrinsic matrix, 2D/3D keypoints, SMPL parameters β, θ, 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.
- Amount: 126,198 images in training and 27,448 images in testing split.

Experiments

- We achieve SOTA results on 3DPW test set and comparable results on Human3.6M validation set. Images in 3DPW are actually captured from a close distance, which is suitable for our model. Human3.6M shows that our translation estimation could also generalize to images from a large distance.

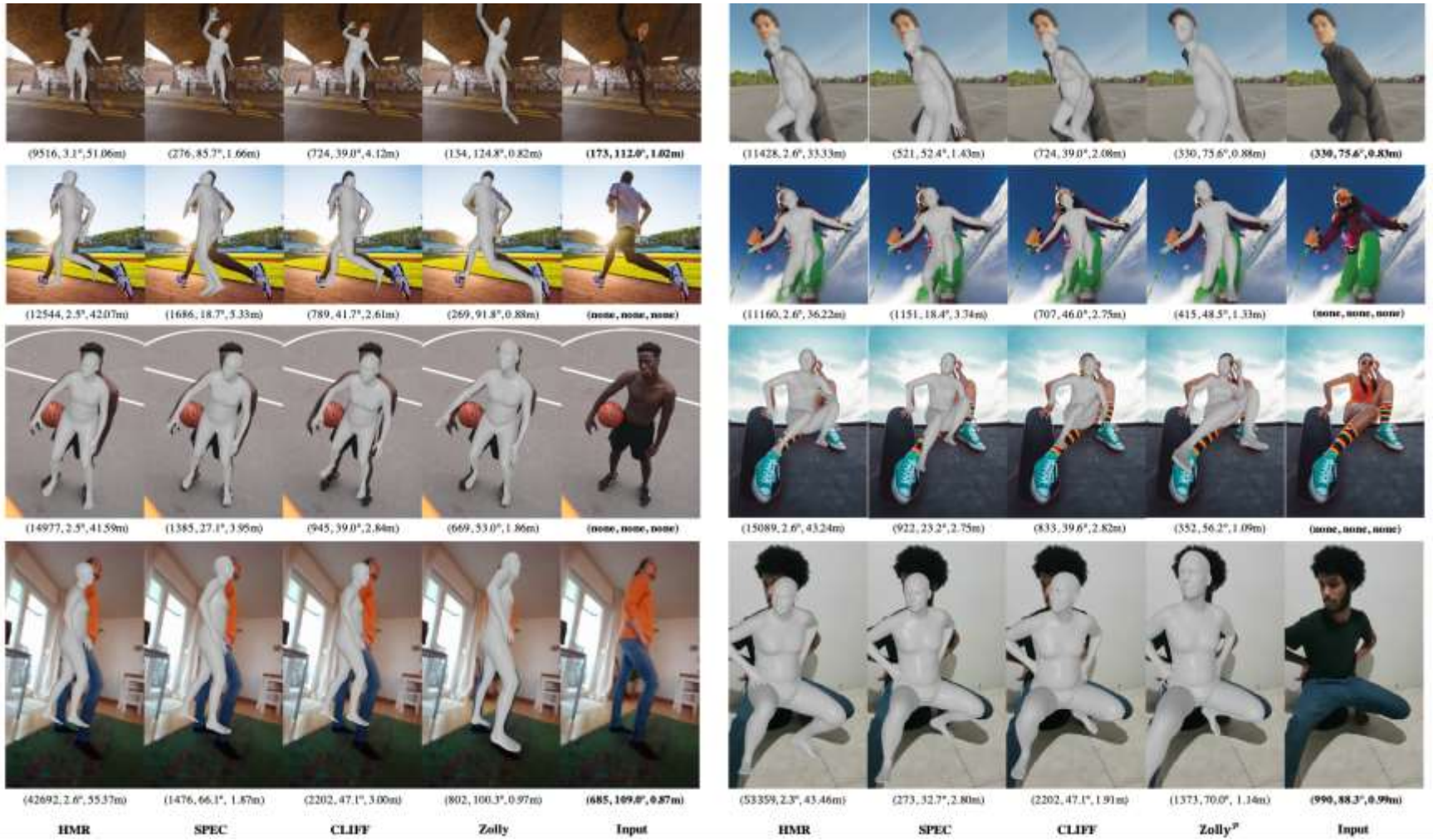
Method	Backbone	3DPW			Human3.6M	
		PA-JPE	MPJPE	PVE	PA-JPE	MPJPE
HMR	Res50	72.6	116.5	-	56.8	88.0
SPEC	Res50	52.7	96.4	-		
CLIFF	HR48	43.0	69.0	81.2	32.7	**47.1**
Zolly	HR48	**39.8**	**65.0**	**76.3**	**32.3**	49.4

Experiments

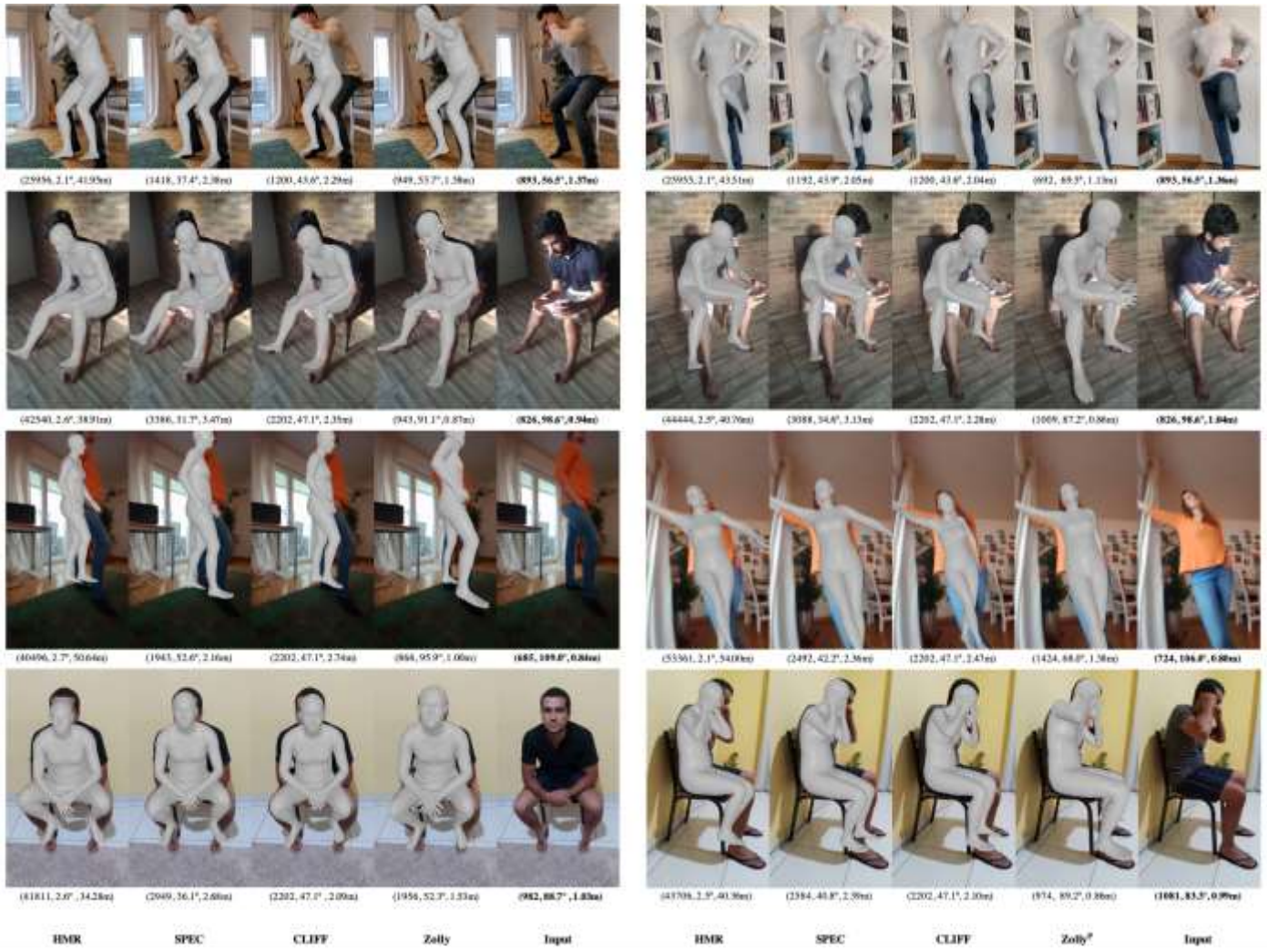
- We achieve SOTA results on PDHuman test set and SPEC-MTP. We chose the protocol with most serve distortion from each dataset here. mIoU here is the mIoU between rendered mask and GT masks, which could be used to describe the ovelay performance vividly.

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**

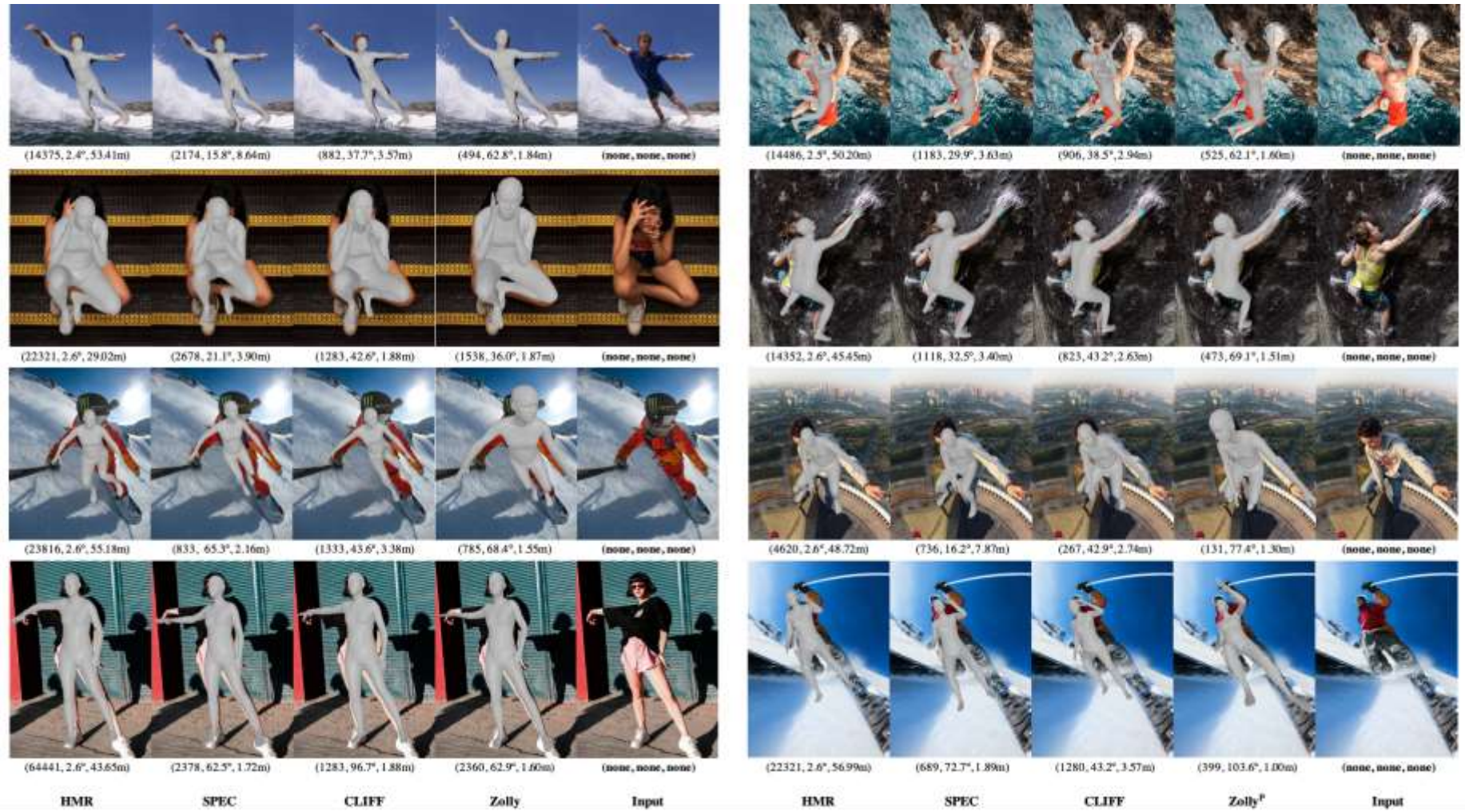
Quality Results



Quality Results on SPEC-MTP



Quality Results on Web Images



Quality Results on Standard Benchmark