

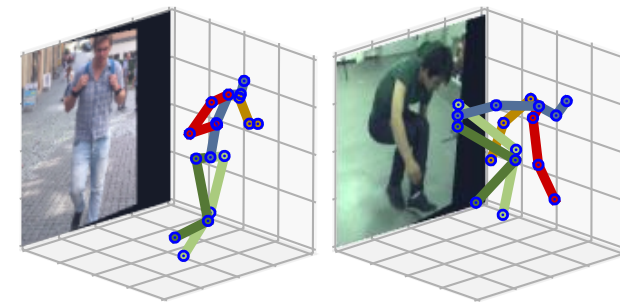
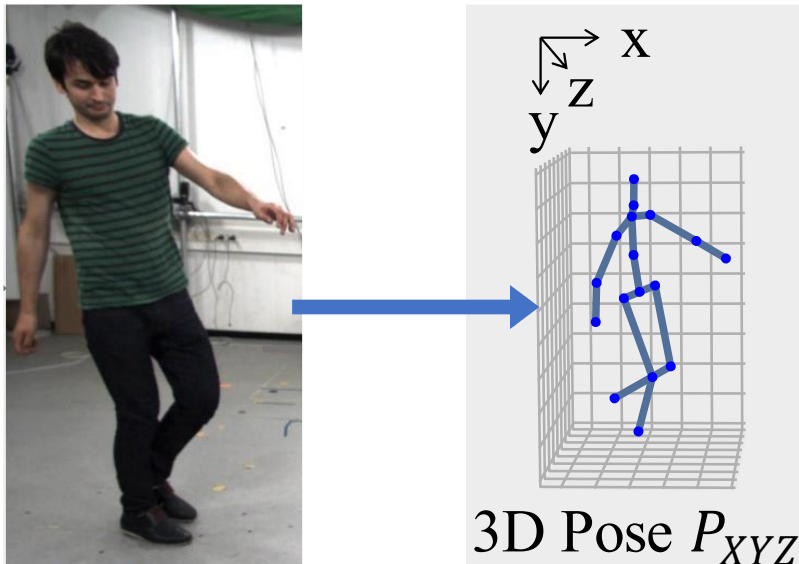
PersPose: 3D Human Pose Estimation with Perspective Encoding and Perspective Rotation

Xiaoyang Hao*, Han Li*
(* equal contribution)

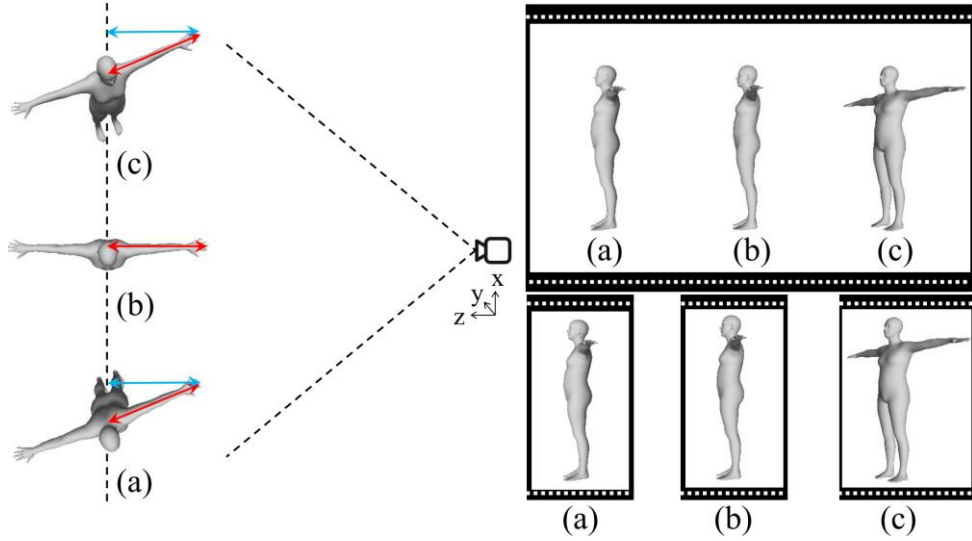


Task

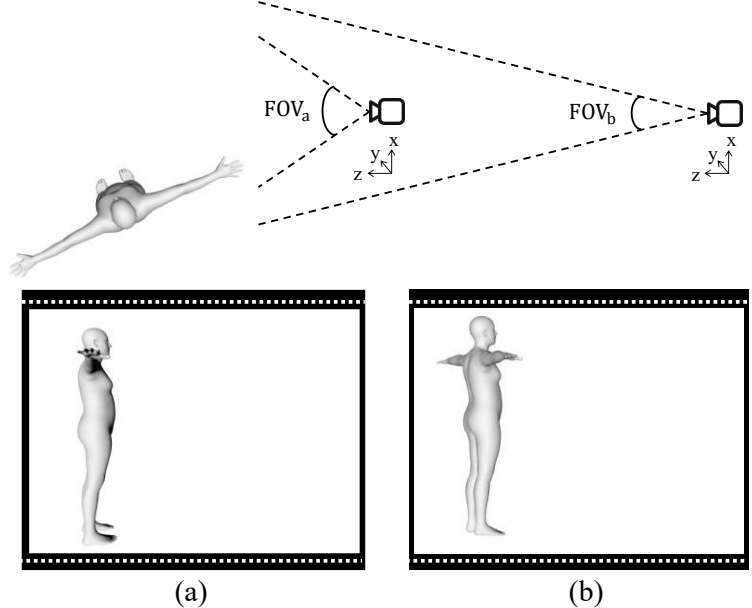
- **3D Human Pose Estimation (HPE):** Estimate the 3D positions of human joints from individual images.
- **Application:** AR/VR and human-computer interaction.
- **Background:** Existing 3D HPE methods typically use **cropped images alone** as input. However, we found that without **camera intrinsics** information, the relative depths of joints cannot be accurately estimated.



Importance of Camera Intrinsics in 3D HPE



- **Case 1:** When **only** input the **cropped image**, joints' relative depths cannot be accurately estimated.



- **Case 2:** Use the full image but without **FOV** info → still inaccurate estimation of joints' relative depths.

Cropping an image is equivalent to capturing a photo using another camera with the same camera extrinsic but with different camera intrinsics and resolution.

$$K^{\text{crop}} = AK = \begin{bmatrix} f^{\text{crop}} & 0 & c_x^{\text{crop}} \\ 0 & f^{\text{crop}} & c_y^{\text{crop}} \\ 0 & 0 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} s & 0 & t_u \\ 0 & s & t_v \\ 0 & 0 & 1 \end{bmatrix}$$

cropping affine

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

original camera intrinsics

$$s = \frac{w^{\text{crop}}}{w^{\text{bbox}}}$$

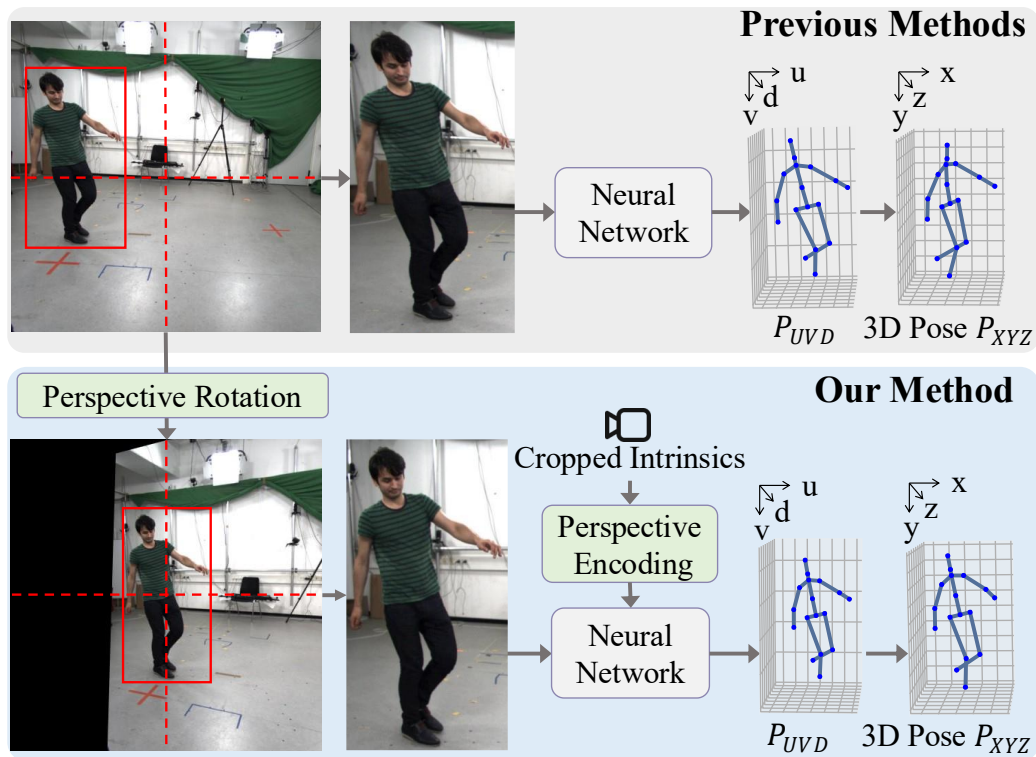
$$t_u = \frac{w^{\text{crop}}}{2} - s c_u^{\text{bbox}}$$

$$t_v = \frac{h^{\text{crop}}}{2} - s c_v^{\text{bbox}}$$

$w^{\text{crop}}, h^{\text{crop}}$: the resolution of the cropped image
 $w^{\text{bbox}}, h^{\text{bbox}}$: the bounding box resolution
 $c_u^{\text{bbox}}, c_v^{\text{bbox}}$: the center coordinates of the bounding box

- **Cropped intrinsics** K^{crop} encapsulate both **cropping** and **FOV** information.

Method



- We design **Perspective Encoding (PE)** to encode cropped intrinsics as a 2D PE map, which is then jointly fed into the CNN with the cropped image.

$$f_{\theta} : (I^{\text{crop}}, f^{\text{crop}}, c_x^{\text{crop}}, c_y^{\text{crop}}) \rightarrow P_{XYZ}$$

PR $\dashrightarrow \tilde{f}_{\theta} : (I^{\text{crop}}, f^{\text{crop}}) \rightarrow P_{XYZ}$

$$K^{\text{crop}} = \begin{bmatrix} f^{\text{crop}} & 0 & c_x^{\text{crop}} \\ 0 & f^{\text{crop}} & c_y^{\text{crop}} \\ 0 & 0 & 1 \end{bmatrix}, I^{\text{crop}} \text{ is the cropped image, } P_{XYZ} \text{ is the 3D pose}$$

- As the human subject can appear anywhere within the original image, the principal point of the cropped images may differ significantly, which complicates model fitting. And the further the human deviates from the image center, the greater the perspective distortions. To address these issues, we propose **Perspective Rotation (PR)** to center the human.

Method

- **Perspective Encoding (PE):** We project virtual sensors onto a fixed reference plane at $z = 1$. For each sensor, the corresponding projected area geometrically represents its unique view frustum. So, we employ this projected area as the encoded representation of the cropped intrinsics.

$$\begin{bmatrix} (1,1) & (1,2) & \cdots & (1,w) \\ (1,2) & (2,2) & \cdots & (2,w) \\ \vdots & \vdots & \ddots & \vdots \\ (h,1) & (h,2) & \cdots & (h,w) \end{bmatrix}$$

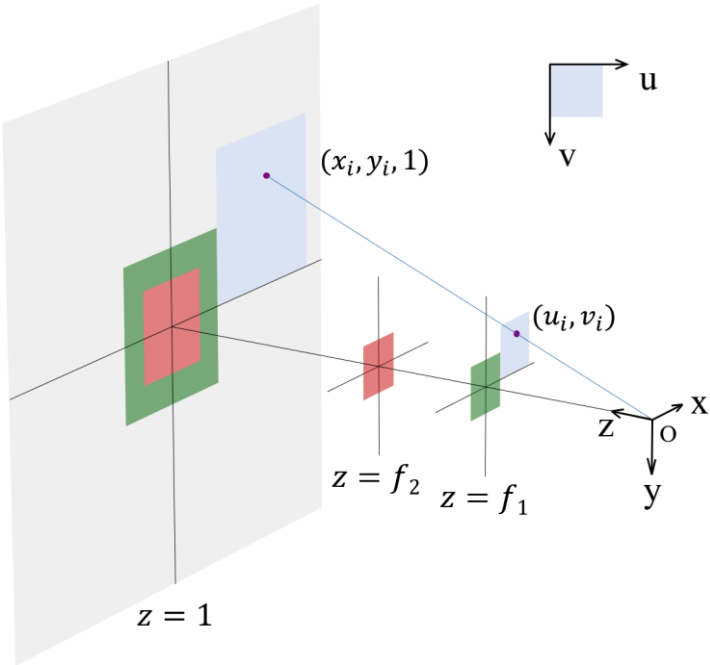
Pixel Coordinate Map M^{uv}

$$\begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = (K^{crop})^{-1} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}$$

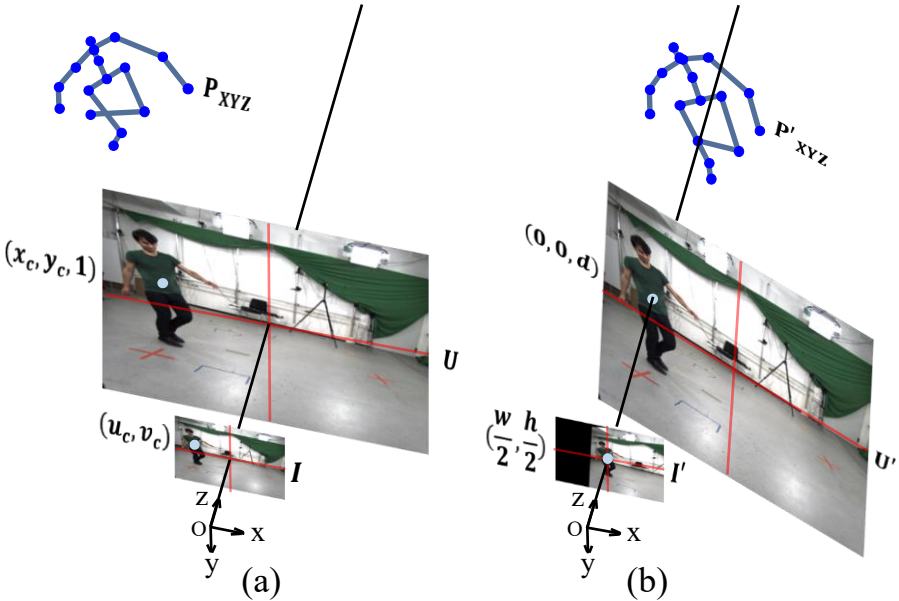
Cropped Intrinsics K^{crop}

$$\begin{bmatrix} (x_1,y_1) & (x_1,y_2) & \cdots & (x_1,y_w) \\ (x_2,y_1) & (x_2,y_2) & \cdots & (x_2,y_w) \\ \vdots & \vdots & \ddots & \vdots \\ (x_h,y_1) & (x_h,y_2) & \cdots & (x_h,y_w) \end{bmatrix}$$

PE Map $M^{xy} (h, w, 2)$

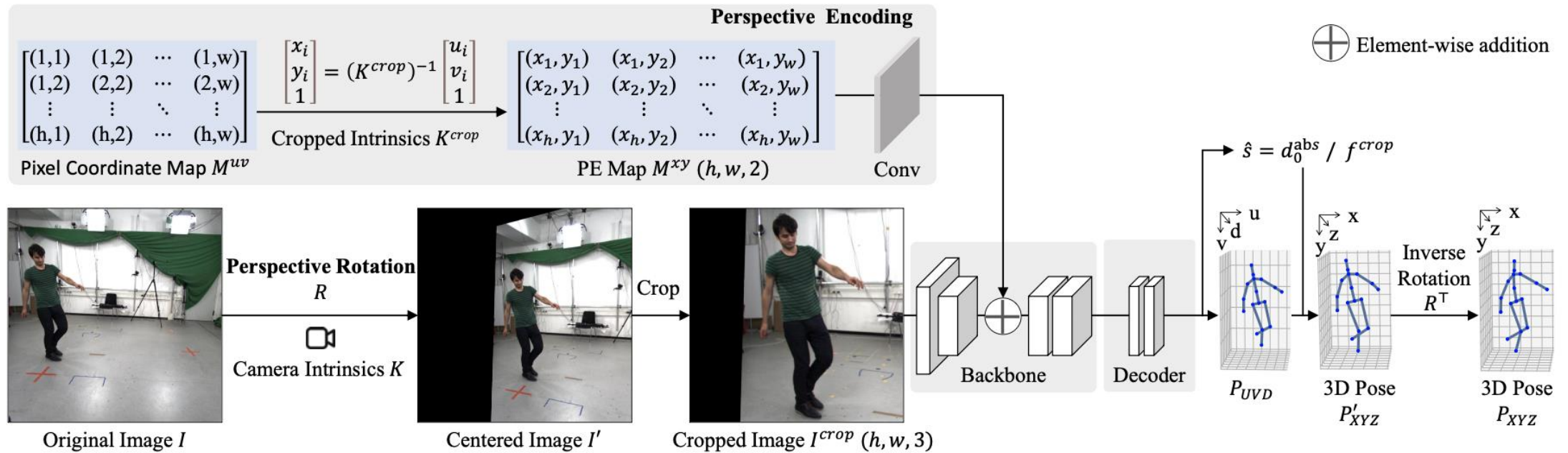


- **Perspective Rotation (PR):** The sensor in (a) captures an image I . An upscaled image U is added on the plane at $z = 1$. A 3D skeleton P_{XYZ} is also added to the scene. P_{XYZ} and U are rotated around the optical center O from (a) to (b), so that the optical axis points to the human. Then, the upscaled image U' in (b) is reprojected onto the sensor to obtain the centered image I' .



Method

➤ Overview of the PersPose



- 1) PR centers the human subject.
- 2) PE encodes the cropped intrinsics into a 2D map M^{xy} .
- 3) PE map and cropped image go through the network to predicts the 2D joint coordinates, the relative depths of joints P_{UVD} , and a scale factor \hat{s} . Then, for a joint index by i , the 3D Pose $P'_{XYZ,i}$ is calculated as:

$$\begin{bmatrix} x'_i \\ y'_i \\ z'_i \end{bmatrix} = d_i^{abs} (K^{crop})^{-1} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}, \quad \text{where } d_i^{abs} = d_i + \hat{s} f^{crop}.$$

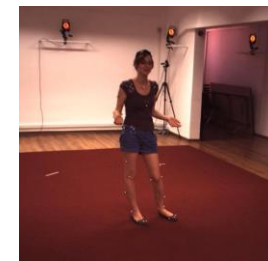
Experimental Results

- PersPose achieves SOTA performance on several datasets.

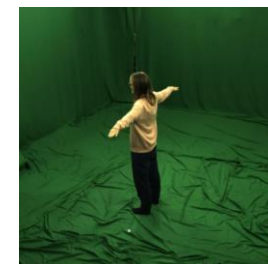
| | | 3DPW | | | Human3.6M | | | MPI-INF-3DHP | | |
|----------------------|---------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|
| | | PA-MPJPE↓ | MPJPE↓ | PVE↓ | PA-MPJPE↓ | MPJPE↓ | PVE↓ | PCK↑ | AUC↑ | MPJPE↓ |
| HMR [8] † | CVPR'18 | 81.3 | 130.0 | 152.7 | 56.8 | 88.0 | 96.1 | 72.9 | 36.5 | 124.2 |
| SPIN [14] † | ICCV'19 | 59.2 | 96.9 | 116.4 | 41.1 | - | - | 76.4 | 37.1 | 105.2 |
| I2L-MeshNet [26] † | ECCV'20 | 57.7 | 93.2 | 110.1 | 41.1 | 55.7 | 65.1 | - | - | - |
| Pose2Mesh [3] † | ECCV'20 | 58.3 | 88.9 | 106.3 | 46.3 | 64.9 | 85.3 | - | - | - |
| Mesh Graphormer [19] | ICCV'21 | 45.6 | 74.7 | 87.7 | 34.5 | 51.2 | - | - | - | - |
| HybrIK [16] ‡ | CVPR'21 | 45.0 | 74.1 | 86.5 | 34.5 | 54.4 | 65.7 | 87.5 | 46.9 | 93.9 |
| CLIFF [18] | ECCV'22 | 43.0 | 69.0 | 81.2 | 32.7 | 47.1 | - | - | - | - |
| FastMETRO [2] | ECCV'22 | 44.6 | 73.5 | 84.1 | 33.7 | 52.2 | - | - | - | - |
| IKOL [42] ‡ | AAAI'23 | 45.5 | 73.3 | 86.4 | - | - | - | 87.9 | 48.1 | 88.8 |
| VirtualMarker [24] ‡ | CVPR'23 | 41.3 | 67.5 | 77.9 | 32.0 | 47.3 | <u>58.0</u> | - | - | - |
| ProPose [4] ‡ | CVPR'23 | 40.6 | 68.3 | 79.4 | <u>29.1</u> | <u>45.7</u> | - | - | - | - |
| PLIKS [29] ‡ | CVPR'23 | 42.8 | 66.9 | 82.6 | 34.7 | 49.3 | - | <u>91.8</u> | <u>52.3</u> | <u>72.3</u> |
| Zolly [35] | ICCV'23 | 39.8 | <u>65.0</u> | <u>76.3</u> | 32.3 | 49.4 | - | - | - | - |
| Gwon et al. [6] | CVPR'24 | 44.3 | 73.2 | 80.3 | - | - | - | - | - | - |
| GLNet-W48 [37] | ECCV'24 | <u>39.5</u> | 66.9 | 77.9 | 29.4 | 48.8 | - | - | - | - |
| PostoMETRO [38] | WACV'25 | 39.8 | 67.7 | 76.8 | - | - | - | - | - | - |
| PersPose ‡ | | 39.1 | 60.1 | 72.4 | 28.3 | 43.0 | 52.7 | 94.0 | 55.2 | 72.1 |



3DPW Sample



Human3.6M Sample

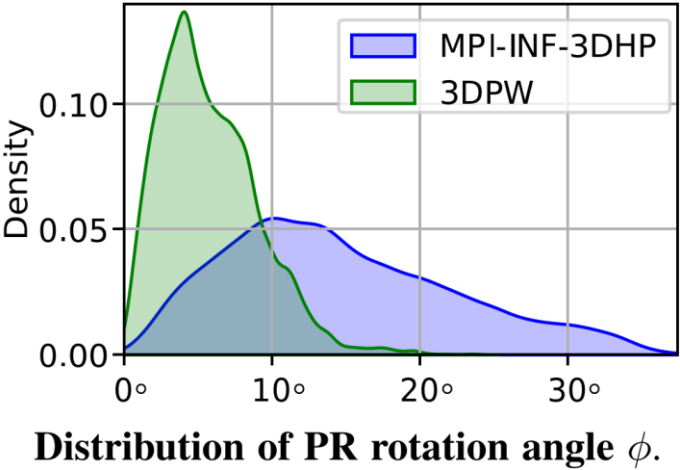


MPI-INF-3DHP Sample

Experimental Results

➤ Ablations show larger gains as perspective complexity increases.

| PR | PE | Dataset | 3DPW | | | MPI-INF-3DHP | | |
|----|----|---------|--------------|-------------|-------------|--------------|-------------|-------------|
| | | | Depth error↓ | PA-MPJPE↓ | MPJPE↓ | Depth error↓ | PA-MPJPE↓ | MPJPE↓ |
| - | - | R | 45.1 | 39.8 | 62.4 | 57.3 | 57.3 | 80.1 |
| - | ✓ | | 44.5 | 39.7 | 62.2 | 53.7 | 56.3 | 76.6 |
| ✓ | ✓ | | 43.8 | 39.1 | 60.1 | 51.0 | 54.4 | 71.9 |
| - | - | R+B | 41.5 | 37.8 | 58.4 | 54.2 | 55.5 | 76.8 |
| - | ✓ | | 41.2 | 37.8 | 58.1 | 51.0 | 55.1 | 73.4 |
| ✓ | ✓ | | 40.0 | 37.3 | 57.2 | 48.6 | 53.7 | 70.2 |



MPI-INF-3DHP spans a wider PR rotation angle range than 3DPW, indicating greater scene–crop perspective change; in such cases, PE/PR help more.

Experimental Results

➤ Visualization.

Crop w/o PR



Crop w/ PR



Our Results

