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

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

\textsuperscript{1} The University of Hong Kong \textsuperscript{2} The University of Adelaide \textsuperscript{3} SenseTime Research \textsuperscript{4} Shanghai AI Laboratory \textsuperscript{5} Zhejiang University

Figure 1: Close-up photography can make it difficult to discern 3D human pose in perspective-distorted images, while state-of-the-art methods often struggle with weak-perspective camera models or inaccurate focal length estimates. Our method overcomes these challenges, accurately recovering 3D human mesh from an approximate distance for fine-grained reconstruction. We could thus get focal length by our proposed approach.

Abstract

As it is hard to calibrate single-view RGB images in the wild, existing 3D human mesh reconstruction (3DHMR) methods either use a constant large focal length or estimate one based on the background environment context, which can not tackle the problem of the torso, limb, hand or face distortion caused by perspective camera projection when the camera is close to the human body. The naive focal length assumptions can harm this task with the incorrectly formulated projection matrices. To solve this, we propose Zolly, the first 3DHMR method focusing on perspective-distorted images. Our approach begins with analyzing the reason for perspective distortion, which we find is mainly caused by the relative location of the human body to the camera center. We propose a new camera model and a novel 2D representation, termed distortion image, which describes the 2D dense distortion scale of the human body. We then estimate the distance from distortion scale features rather than environment context features. Afterwards, We integrate the distortion feature with image features to reconstruct the body mesh. To formulate the correct projection matrix and locate the human body position, we simultaneously use perspective and weak-perspective projection loss. Since existing datasets could not handle this task, we propose the first synthetic dataset PDHuman and extend two real-world datasets tailored for this task, all containing perspective-distorted human images. Extensive experiments show that Zolly outperforms existing state-of-the-art methods on both perspective-distorted datasets and the standard benchmark (3DPW). Code and dataset will be released at \url{https://wenjiawang0312.github.io/projects/zolly/}.

1. Introduction

Human pose and shape estimation from single-view RGB images is a long-standing research area of computer vision, as the reconstructed motion and mesh could empower various human-centered downstream applications like 3D animations, robotics, or AR/VR development. Previous works \cite{19, 26, 23, 24, 41, 23, 27, 20, 55} formulate the problem under the assumption that the reconstructed people are far away from the camera, thus the torso and limb distortion caused by the perspective projection can be neglected.
However, perspective distortion in close-up images is common in real-life scenarios, such as photographs of athletes/actors in sports events/films or selfies taken for social media. In such images, distortions are usually caused by aerial photography, overhead beat, or large depth variance among torsos and limbs, resulting in depth ambiguity in single-view RGB images, which makes it a big challenge to recover human pose and shape (See Fig. 1).

Previous methods typically assume a large fixed focal length [19, 26, 23, 24] or estimate a focal length [25] using pre-trained networks and calculate the translation from the estimated focal length. These settings are appropriate when people are far from the cameras, where the depth variance of the human body is negligible compared to the distance to the camera. However, these methods are inappropriate for handling scenarios in which human bodies are perspective distorted. Overestimating the focal length could lead to joint angle ambiguity or harm joint rotation learning. Several methodologies for pose estimation, as proposed by previous works [22, 29], assume a large field-of-view (FoV) angle. However, these methods may not show significant improvement when the focus is solely on non-distorted human images, as they often lack a conditioning for depth variance when the camera zooms in or out. Inaccuracies in estimating the depth variance with respect to translation can adversely impact re-projection loss, leading to erroneous results, as illustrated in Fig. 1. Actually, a correctly estimated distance and focal length also help with 2D alignment, which will be useful in downstream tasks.

To address the challenge of perspective distortion in close-up images, showing respect to Hitchcock’s dolly zoom shot, we introduce Zolly (Zoom fOcal Length correctly) for perspective-distorted human mesh reconstruction. Our method utilizes 2D human distortion features to estimate the real-world distance to the camera center, enabling the reconstruction of the 3D human mesh in perspective-distorted images. The framework comprises of two parts: a translation estimation module for estimating the z-axis distance of the human body from the camera center, and a mesh reconstruction module for reconstructing 3D vertex coordinates in camera space. Additionally, we introduce a hybrid loss function that combines both perspective and weak perspective projection to boost performance.

Inspired by the iconic dolly-zoom shot [39] (also known as zolly shot), which creatively combines camera movement and zooming to create a distorted perspective and sense of unease, we propose a translation estimation module for the perspective-distorted 3DHMR task. This module highlights how the relative position of the human body to the camera affects the perspective distortion in images. Based on this insight, we introduce the distorted image as a new representation to capture the 2D shrinking or dilation scales of each pixel. Our translation network utilizes distortion and IUV images to accurately estimate z-axis translation, overcoming the limitations of traditional methods that rely on environmental information. IUV image could help eliminate the 2D shift and scale information in distorted images and represent 2D dense position information. For mesh reconstruction, we lift the 2D position feature to the 3D vertex position feature and sample the by-vertex distortion feature to regress 3D vertex coordinates. We use perspective projection to supervise correctly and weak-perspective projection to locate the 2D human body position in the image and help to calculate our focal length.

In summary, our contributions are as follows:

1. We analyze the state-of-the-art (SOTA) 3DHMR methods and propose a novel approach tailored to the perspective-distorted 3DHMR task.

2. We propose a novel learning-based method to tackle the perspective-distorted 3DHMR task without relying on extra camera information. The core of our method is a newly designed representation, termed distortion image, and a hybrid projection supervision that make use of both perspective and weak-perspective projection. (3) We build the first large-scale synthetic dataset PDHuman for the perspective-distorted 3DHMR task, with high-quality SMPL ground truth and camera parameters. To evaluate the performance on real images, we prepare two real-world benchmark datasets, SPEC-MTP [25] and HuMMan [6], which contain perspective-distorted images with well-fitted SMPL parameters and camera parameters.
camera center, resulting in errors in the reconstructed 3D system. When the person is close to the camera model is formed as a weak-perspective camera, with a constant focal length of \( f \). Camera system analysis. In weak-perspective projection, the inner depth variance is ignored in the human body, which means this projection model views the human body as a planar object without thickness. Thus the projection matrix should be as follows:

\[
\begin{bmatrix}
  f & 0 & -x_c \\
  0 & f & -y_c \\
  0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
  x' \\
  y' \\
  1
\end{bmatrix}
= \begin{bmatrix}
  f(x + T_x) \\
  f(y + T_y) \\
  f(z + T_z)
\end{bmatrix},
\]

\[ z = 0, \] (1)

where \( f \) refers to the focal length in NDC (Normalized Device Coordinate) space, \( x, y, z \) refers to a vertex point on human body mesh and \( T_x, T_y, T_z \) refers to pelvis translation. The weak-perspective camera parameters \((s, t_x, t_y)\), which represent 2D orthographic transformation, could be used to approximate the projection:

\[
\begin{bmatrix}
  f(x + T_x)/T_z \\
  f(y + T_y)/T_z
\end{bmatrix} = \begin{bmatrix}
  s(x + t_x) \\
  s(y + t_y)
\end{bmatrix}.
\] (2)

2. Related Work

Mainstream 3D human mesh reconstruction methods. 3D human pose estimation from a single RGB image is essentially an ill-posed problem. To obtain more realistic and manipulable human bodies, a parametric body model SMPL [33] was proposed, which uses 3D rotation representation to model human joint motions with defined LBS weights. To reconstruct human mesh from RGB images, there exist two mainstream pipelines: optimization-based methods and learning-based methods.

Optimization-based methods [5, 14] directly fit the body model parameters to 2D evidence via gradient back-propagation in an iterative manner. Learning-based approaches [19, 36, 52, 24] leverage a deep neural network to regress the human body model parameters or 3D coordinates of the human mesh, which can be further divided into model-based and model-free methods. Inspired by 3D mesh reconstruction tasks [10, 12, 54, 48, 47, 16], Model-based methods works [19, 25, 24, 28, 51, 50] utilize SMPL parameters to recover the human pose and shape. The milestone method HMR [19] takes it as a direct regression task.

Model-free works [27, 30, 8, 11] directly reconstructing 3D meshes from single view images.

Human mesh reconstruction with specific camera systems. In the previous trend led by HMR [19], the intrinsic camera model is formed as a weak-perspective camera, with a constant focal length of 5,000 pixels. However, this assumption does not hold well when the person is close to the camera center, resulting in errors in the reconstructed 3D shape and pose. To address this issue, several recent works such as BeyondWeak [22], SPEC [25], and CLIFF [29] have proposed different camera system assumptions. SPEC predicts camera parameters (pitch, yaw, and FoV) from a single-view image, but its asymmetric Softargmax-\(L_2\) loss tends to overestimate focal length and translation, which is not suitable for distorted images. Moreover, SPEC regresses camera parameters through environmental information, which can sometimes be meaningless when the background lacks geometry information. CLIFF focuses on joint rotation variance caused by horizontal shift but has not conditioned the distance from the human body to the camera. CLIFF, following BeyondWeak [22], uses the diagonal length of the image as the focal length, which is not a close assumption for distorted problems since the focal length can be easily adjusted during image capture.

Compared to these methods, our framework estimates the z-axis translation from 2D human distortion features, and obtains a more accurate focal length from the estimated translation, leading to much better reconstruction accuracy on distorted images. See a comparison of the camera models in Fig. 3. In the Sup. Mat., we quantitatively demonstrate the bad re-projection influence caused by a wrongly formulated projection matrix.

3. Methodology

In this section, we first review the formulation of previous camera systems and then present our camera system customized for distorted images in Sec. 3.1. Sec. 3.2 presents our network architecture with two key components: (i) translation estimation module and (ii) mesh reconstruction module. Subsequently, we explain the proposed hybrid re-projection loss functions for distorted human mesh reconstruction in Sec. 3.3.

3.1. Preliminary

Camera system analysis. In weak-perspective projection, the inner depth variance is ignored in the human body, which means this projection model views the human body as a planar object without thickness. Thus the projection matrix should be as follows:

\[
\begin{bmatrix}
  f & 0 & -x_c \\
  0 & f & -y_c \\
  0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
  x' \\
  y' \\
  1
\end{bmatrix}
= \begin{bmatrix}
  f(x + T_x) \\
  f(y + T_y) \\
  f(z + T_z)
\end{bmatrix},
\]

\[ z = 0, \] (1)

where \( f \) refers to the focal length in NDC (Normalized Device Coordinate) space, \( x, y, z \) refers to a vertex point on human body mesh and \( T_x, T_y, T_z \) refers to pelvis translation. The weak-perspective camera parameters \((s, t_x, t_y)\), which represent 2D orthographic transformation, could be used to approximate the projection:

\[
\begin{bmatrix}
  f(x + T_x)/T_z \\
  f(y + T_y)/T_z
\end{bmatrix} = \begin{bmatrix}
  s(x + t_x) \\
  s(y + t_y)
\end{bmatrix}.
\] (2)
Finally, we can get:
\[
(3)
\]
However, the perspective projection actually is:
\[
(4)
\]
From Eq. (4), if \( z \) gets smaller, the projected \( x_{2D}, y_{2D} \) will be larger. This phenomenon causes the closer points on the close-up photographed image to dilate and the farther points to shrink. Thus, the 3D translation results in pixel-level distortion on the limbs, torso, or faces in the 2D projected image. Usually, when the human body is farther than 5 m, the distortion is subtle. Under these circumstances, a weak-perspective projection could be used.

Following the weak-perspective assumption, we take the \( f = s \times T_z \) (\( f \) is in NDC space) as an approximation. \( T_z \) is the z-axis translation of the pelvis, which could be viewed as a mean translation of the whole human body. The difference compared to previous methods [19, 25, 29] is that we first estimate the body translation and then calculate the focal length. So we still need to estimate weak-perspective camera parameters \((s, t_x, t_y)\) to compute the focal length and obtain the 2D location in the image. Following SPEC [25], we get the \( T_x, T_y \) in the full image from the \( t_x, t_y \) by affine transformation using the bounding box.

**Distortion image.** As described in Sec. 3.1, our approach projects the 3D translations of human body points into 2D images where the limbs dilate or shrink. We adopt the \( x-y \) plane of the pelvis as the reference plane representing a ‘scale equals 1’ plane. When the human is distant from the camera, it can be approximated as a zero-thickness plane, where all distortion scales are 1. And as shown in Eq. (4), distortion scales are inversely proportional to the \( z \)-distance from the camera when the body is closer. To quantify limb distortion caused by perspective projection, we introduce a distortion image \( I_d \), where \( I_d = T_z/I_{Depth} \), where \( I_{Depth} \) represents a depth image. The distorted image and its pixel value enable a visual and numerical representation of the limb dilation or shrinkage caused by the perspective camera. For instance, when the pelvis is fixed, a finger can appear twice as dilated when \( z \) of the finger reaches from 1 m to 0.5 m. See Fig. 5 for the demonstration of distortion image.

### 3.2. Network Structure

Given a monocular image, Zolly applies an off-the-shelf Convolution Neural Network, e.g. [15, 40], as an image encoder; the output multi-level features can be used as an input for the translation module and the mesh reconstruction module that we describe next.

**Translation module.** As shown in Fig. 4, we estimate the distortion image \( I_d \) and IUV image \( I_{UV} \) with an FPN [31] structure. There are two main advantages for this setting. Firstly, we can distillate the geometry information on the human body without background context. Another advantage is that this dense correspondence predicting task is easy to train. As noted in Sec. 3.1, the distortion type corresponds to one certain translation, and the distortion is determined when the image was captured, whether or not...
cropped or rotated afterwards. So we further warp the distorted image into the continuous UV space \([52]\) to eliminate the 2D scale, shift, and rotation. We treat \(T_z\) as a learnable embedding. A \(1 \times 1\) convolution is first applied to up-sample the channels of the warped distortion image. Then cross-attention \([44]\) is performed between the warped distortion feature and the z-axis embedding, with a fully connected layer to output \(T_z\). Note that we use sigmoid then \(\times 10\) to restrict \(T_z\) to be between 0 and 10.n. Following SPEC \([25]\), we get \(T_x\) and \(T_y\) by applying affine transform on the estimated \(t_x, t_y\) with ground-truth bounding boxes (See Sup. Mat. for more details). The loss function of the translation module is formulated as follows:

\[
L_{Trans} = \lambda_{IUV} L_{IUV}^1 + \lambda_d L_d^2 + \lambda_z L_z^1,
\]  

where \(L_{IUV}^1\) is the \(L2\) loss of the IUV image, \(L_d^2\) is the \(L2\) loss of the distortion image, and \(L_z^1\) is the \(L1\) loss of z-axis translation.

**Mesh reconstruction module.** Different from previous methods that use graph convolution \([27]\) or transformers \([8, 30]\) for building long-range dependence among different vertices, we adopt a light-weight MLP-Mixer \([42]\) structure to model the attention among different vertices, followed by a fully connected layer that lifts per-vertex position features \(F_{vp}\) from the spatial feature \(F_{grid}\) which was used to predict \(I_{IUV}\).

As illustrated in Fig. 4, since the distortion feature has already been warped into UV space, we could easily sample the per-vertex distortion feature \(F_{vd}\) from the warped distortion feature \(F_d\) by pre-defined Vertex UV coordinates \(V_{uv}\) \([52]\). We concatenate \(F_{vd}\) with \(F_{vp}\) and use fully connected layers to predict the coordinates of a coarse mesh of the body that is composed of 431 vertices. The coarse mesh is up-sampled using two fully connected layers, resulting in an intermediate mesh with 1,723 vertices and a full mesh with 6,890 vertices. 3D joint coordinates are obtained using a joint regression matrix provided by the SMPL \([33]\) body model. The total loss for the mesh reconstruction module is:

\[
L_{Mesh} = \lambda_{J3D} L_{J3D}^1 + \lambda_{J2D} L_{J2D}^1 + \lambda_{LW} L_{LW}^1 + \lambda_V (L_V^{1,\mu} + L_V^{1,\nu} + L_V^{1,\lambda}),
\]  

where \(L_{J3D}^1\) is \(L1\) loss of 3D joints, \(L_{J2D}^1\), \(L_{LW}^1\), and \(L_V^1\) are \(L1\) loss of coarse, intermediate vertices, and full vertices respectively. \(L_{J2D}^1\) and \(L_{LW}^1\) represent loss of perspective and weak-perspective re-projected 2D joints, and will be further illustrated in Sec. 3.3.

### 3.3. Hybrid Re-projection Supervision

Most existing methods \([19, 29, 8]\) usually use a pre-defined focal length \(f\). SPEC \([25]\) train a CamCalib network to estimate the focal length. Then, z-axis translation \(T_z\) can be calculated by \(T_z = 2f/hs\). On the contrary, as illustrated in Eq. (3), we aim to get the focal length \(f\) by directly predicting the orthographic scale \(s\) and z-axis translation \(T_z\). Following HMR \([19]\), we still use the weak-perspective projection besides perspective projection.

**Weak-perspective re-projection.** For weak-perspective projection loss, we follow HMR \([19]\), use focal length \(f_W\) as 5,000 pixels, and thus formulate the weak-perspective intrinsic matrix and translation separately as:

\[
K_W = \begin{bmatrix} f_W & h/2 & t_x \\ f_W & h/2 & t_y \\ 1 & 1 & 2f_W/sh \end{bmatrix},
\]

Then we project the 3D joints \(\hat{J}_{3D}\) and measure the difference with 2D keypoints in image coordinates as:

\[
\hat{J}_{3D} = K_W (\hat{J}_{3D}^0 + T_W),
\]

\[
L_{2D}^W = \sum_{i=1}^{N_j} \frac{1}{d_{j[i]}} \| J_{2D}^W[i] - J_{2D}^i \|_F^2,
\]

where \(\hat{J}_{3D}^0\) means we detach the gradient from the body model joints in weak-perspective projection. This means we only update the weak-perspective camera \((s, t_x, t_y)\) and do not want this wrong projection to harm the body pose gradient flow. \((s, t_x, t_y)\) are mainly used to locate the human body’s position in image coordinates and compute the focal length \(f_P\). For better position alignment, we divide a distortion weight \(d_{j[i]}\), which is sampled from distortion image \(I_d\) by \(J_{2D}^0[i]\) for every joint. This forces the dilated limbs to get a smaller weight while the shrunk limbs get a bigger weight.

**Perspective re-projection.** Perspective re-projection is mainly used to supervise pose or mesh reconstruction with the correct projection matrix. Firstly, we have 3D joints \(\hat{J}_{3D} = \delta_{reg} V\). We use ground-truth focal length \(f_P\) to stabilize the training. For samples without ground-truth focal length, we will use a focal length of 1,000 pixels for 224×224 images. This will make the translation range approximately from 5 to 10 meters. During inference, according to Eq. (4), we compute the focal length in screen space for perspective projection by \(f_P = shT_z/2\) pixels, where \(h\) represents cropped image height, equals 224 pixels in our setting. Thus we can formulate the perspective intrinsic matrix \(K_P\) and projected 2D joints \(\hat{J}_{2D}^P\) as:

\[
K_P = \begin{bmatrix} f_P & H/2 & t_x \\ f_P & H/2 & t_y \\ 1 & 1 & 2f_P/sh \end{bmatrix}, \hat{J}_{2D}^P = K_P (\hat{J}_{3D} + T_{P}^0),
\]

where \(T_{P}^0\) is the translation estimated by translation head in Sec. 3.2. We detach it as well to avoid the alignment conflicting of two re-projection. We project the 3D joints \(\hat{J}_{3D}\)
Figure 5: Demonstration of distortion image. The three columns from left to right are our PDHuman dataset, HuMMan dataset, and SPEC-MTP dataset, respectively. The value from the arrow (yellow) indicates the distortion scale of the pixel.

and measure the difference with the original 2D keypoints in the image coordinates before cropped.

4. Experiments

4.1. Datasets

PDHuman. Despite perspective distortion being a common problem, no existing public dataset is specifically designed for this task. Inspired by recent synthetic datasets [49, 7, 3, 35], we introduce a synthetic dataset named PDHuman. The dataset contains 126,198 images in the training split and 27,448 images in the testing split, with annotations including camera intrinsic matrix, 2D/3D keypoints, SMPL parameters ($\theta$, $\beta$), and translation for each image. The testing split is further divided into 5 protocols by the max distortion scale of each image sample. We define the max distortion scale for each sample as $\tau$; this value will be used in splitting protocols.

We use 630 human models from RenderPeople [2] and 1,710 body pose sequences from Mixamo [1], with 500 HDRi images with various lighting conditions as backgrounds. We use the dolly-zoom effect to generate random camera extrinsic and intrinsic matrices with random rotations, translations, and focal lengths. The distance from the human body to the cameras is set from 0.5m to 10m, so our dataset contains severely distorted, slightly distorted, and nearly non-distorted images. Then we use Blender [4] to render the RGB images. See Fig. 5 for brief demonstration. For detailed rendering procedures and more image demonstrations, please refer to Sup. Mat.

SPEC-MTP and HuMMan Datasets. SPEC-MTP dataset [25] is proposed to test human pose reconstruction in world coordinates. It includes many close-up shots, mostly taken from below or overhead views, leading to images with distorted human bodies. HuMMan dataset [6] is captured by multi-view RGBD cameras and has accurate ground truth because the SMPL parameters are fitted based on 3D keypoints and point clouds. HuMMan also contains images with distorted human bodies since the actors were close to the cameras, all less than 3 meters away. In our paper, we extend both datasets into real-world perspective-distorted datasets. SPEC-MTP is used only for testing. For HuMMan, we split it into training and testing parts. When testing, we divide these two datasets into three protocols based on their maximum distortion scale $\tau$. See Fig. 5 for brief demonstration.

Non-distorted Datasets. For non-distorted datasets, we use Human3.6M [17], COCO [32], MPI-INF-3DHP [34] and LSPET [18] as our training data. Following [30, 8], we also report the results fine-tuned on 3DPW [45] training data.

4.2. Evaluation Metrics

To measure the accuracy of reconstructed human mesh, we follow the previous works [19, 25, 29] by adopting MPJPE (Mean Per Joint Position Error), PA-MPJPE (Procrustes Analysis Mean Per Joint Position Error) and PVE (Per Vertex Error) as our 3D evaluation metrics. They all measure the Euclidean distances of 3D points or vertices between the predictions and ground truth in millimeters (mm).

To measure the re-projection results in perspective distorted datasets such as PDHuman, SPEC-MTP and HuMMan, we leverage metrics widely used in segmentation tasks, MeanIoU [13] as our 2D metric. We both report foreground and background MeanIoU marked as mIoU and body part MeanIoU marked as $P$-mIoU. We use the 24-part vertex split provided by official SMPL [33] for body part segmentation. During the evaluation, for weak-perspective methods like HMR, we will render the predicted segmentation masks with a focal length of 5,000 pixels. And we use the corresponding focal length on methods with specific camera models, such as SPEC [25], CLIFF [29], and proposed Zolly.

4.3. Implementation Details

Unless specified, we use ResNet-50 [15] and HRNet-w48 [40] backbones for model-free Zolly. We also design a model-based variant, Zolly$^p$ ($p$ stands for parametric), by changing the mesh reconstruction module to a model-based pose and shape estimation module. The details of Zolly$^p$ can be found in the Sup. Mat. All backbones are initialized by COCO [32] key-point dataset pre-trained models. We use Adam [21] optimizer with a fixed learning rate of $2e^{-4}$. All experiments of Zolly are conducted on 8 A100 GPUs for around 160 epochs, 14~18 hours. Our training pipeline was built based on MHHuman3D [9] code base. For samples with ground-truth focal length and translations, we render IUV and distortion images online during the training by PyTorch3D [38].
Table 1: Results of SOTA methods on PDHuman, SPEC-MTP [25] and HuMMan [6] datasets. Here we report the largest distortion protocol. R50 terms ResNet-50 [15], and H48 terms HRNet-w48 [40] here.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Backbone w. 3DPW</th>
<th>Metrics</th>
<th>PA-MPJPE</th>
<th>MPJPE↓</th>
<th>PVE↓</th>
<th>mIoU↓</th>
<th>P-mIoU↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zolly (H48)</td>
<td>✓</td>
<td>HybrIK [28]</td>
<td>48.6</td>
<td>80.0</td>
<td>94.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (R50)</td>
<td>×</td>
<td>ResNet-34</td>
<td>44.5</td>
<td>74.4</td>
<td>96.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (H48)</td>
<td>✓</td>
<td>ResNet-50</td>
<td>72.6</td>
<td>116.5</td>
<td>90.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GraphCMR (H48)</td>
<td>✓</td>
<td>ResNet-64</td>
<td>52.7</td>
<td>96.4</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GraphCMR (R50)</td>
<td>×</td>
<td>ResNet-50</td>
<td>52.3</td>
<td>82.9</td>
<td>99.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (H48)</td>
<td>✓</td>
<td>ResNet-50</td>
<td>48.3</td>
<td>77.9</td>
<td>90.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (R50)</td>
<td>×</td>
<td>ResNet-50</td>
<td>45.7</td>
<td>72.0</td>
<td>85.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARE [34]</td>
<td>✓</td>
<td>ResNet-w32</td>
<td>46.5</td>
<td>74.5</td>
<td>88.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARE [24]</td>
<td>×</td>
<td>ResNet-50</td>
<td>43.0</td>
<td>69.0</td>
<td>81.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HybrIK [28]</td>
<td>✓</td>
<td>ResNet-w64</td>
<td>45.6</td>
<td>74.7</td>
<td>87.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HybrIK [28]</td>
<td>×</td>
<td>ResNet-64</td>
<td>44.6</td>
<td>73.5</td>
<td>84.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (R50)</td>
<td>×</td>
<td>ResNet-50</td>
<td>48.9</td>
<td>80.0</td>
<td>92.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (H48)</td>
<td>✓</td>
<td>ResNet-50</td>
<td>49.2</td>
<td>79.6</td>
<td>92.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (R50)</td>
<td>×</td>
<td>ResNet-50</td>
<td>44.1</td>
<td>72.5</td>
<td>84.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (H48)</td>
<td>✓</td>
<td>ResNet-w48</td>
<td>47.9</td>
<td>76.2</td>
<td>89.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zolly (R50)</td>
<td>×</td>
<td>ResNet-w48</td>
<td>39.8</td>
<td>65.0</td>
<td>76.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results of SOTA methods on 3DPW. Zolly\textsuperscript{P} terms our parametric-based variant. \textquoteleft w/o PD\textquoteright means trained without the proposed distorted dataset. \textquoteleft w/ f\textquoteright means trained with ground-truth focal length if provided.

Results on PDHuman, SPEC-MTP, and HuMMan. We report PA-MPJPE, MPJPE, PVE, mIoU, and P-mIoU on these three datasets. For model-based methods, we compare with HMR [19], SPEC [25], CLIFF [29], PARE [24]. We report PA-MPJPE, MPJPE, PVE, mIoU, and P-mIoU on these three datasets. For model-based methods, we compare with HMR [19], SPEC [25], CLIFF [29], PARE [24]. We compare model-free methods with GraphCMR [27] and FastMETRO [8]. From Tab. 1, we can see that SPEC [25] performs poorly on these distorted datasets. This is mainly due to their wrong focal length assumption, which has negative rather than positive effects on their supervision. (Note that our re-implemented SPEC has higher performance than the official code, see in Sup. Mat.) CLIFF performs well on SPEC-MTP, while badly on PDHuman. Because their focal length assumption is about 53° for 16:9 images, close to SPEC-MTP images. Although HMR-\textit{f} is trained with the same focal length as Zolly, it improves little compared to HMR since they have not encoded the distortion or distance feature into their network. Zolly-H48 outperforms SOTA methods on most metrics, especially the 3D ones. Some 2D re-projection metrics, e.g. mIoU and P-mIoU, of Zolly-H48 are lower than Zolly\textsuperscript{P}-R50 version. We conjecture that model-based methods have better reconstructed shapes. Please refer to Fig. 6 for qualitative results. More qualitative results and failure cases can be found in Sup. Mat.

4.4. Main Results

Results on PDHuman, SPEC-MTP, and HuMMan. We report PA-MPJPE, MPJPE, PVE, mIoU, and P-mIoU on these three datasets. For model-based methods, we compare with HMR [19], SPEC [25], CLIFF [29], PARE [24]. We compare model-free methods with GraphCMR [27] and FastMETRO [8]. From Tab. 1, we can see that SPEC [25] performs poorly on these distorted datasets. This is mainly due to their wrong focal length assumption, which has negative rather than positive effects on their supervision. (Note that our re-implemented SPEC has higher performance than the official code, see in Sup. Mat.) CLIFF performs well on SPEC-MTP, while badly on PDHuman. Because their focal length assumption is about 53° for 16:9 images, close to SPEC-MTP images. Although HMR-\textit{f} is trained with the same focal length as Zolly, it improves little compared to HMR since they have not encoded the distortion or distance feature into their network. Zolly-H48 outperforms SOTA methods on most metrics, especially the 3D ones. Some 2D re-projection metrics, e.g. mIoU and P-mIoU, of Zolly-H48 are lower than Zolly\textsuperscript{P}-R50 version. We conjecture that model-based methods have better reconstructed shapes. Please refer to Fig. 6 for qualitative results. More qualitative results and failure cases can be found in Sup. Mat.
Figure 6: Qualitative results of SOTA methods. Besides Zolly, we visualize the results of three methods with specific camera models: HMR [19], SPEC [25], CLIFF [29]. Zolly terms our model-based variance. We show results come from different data sources. Row 1: PDHuman test. Row 2, 3: web images. Row 4: SPEC-MTP. The number under each image represents predicted/ground-truth focal length, FoV angle, and translation.

Results on Human3.6M: During training, we get the ground-truth focal length and translation from Human3.6M [17] training set for our supervision. When evaluating Human3.6M, we follow HybrIK [28] by using SMPL joints as the ground truth for evaluation. As shown in Tab. 3, our method performs well on Human3.6M through it is not a perspective-distorted dataset. Zolly-H48 achieves the best result on the PA-MPJPE metric and achieves comparable results on the MPJPE metric. CLIFF achieves the best results on MPJPE while they also need ground-truth bounding boxes during testing.

4.5. Ablation study

Ablation on training settings on the standard benchmark 3DPW. In Table 4, we present the results of our ablation study on the standard 3DPW benchmark [45], where we investigate the impact of different training settings on the performance of our method. By controlling two different variables, we show that introducing perspective-distorted datasets and fine-tuning with ground-truth focal length both lead to a slight improvement in performance. Notably, our method Zolly-H48 still outperforms the current state-of-the-art methods even without using perspective-distorted data or ground-truth focal length.

This study evaluates the effectiveness of the distortion feature and the hybrid re-projection loss function. The evaluation is conducted on the PDHuman ($\tau = 3.0$), as this ex-
Figure 7: Qualitative results for 3DPW. Zolly achieves good alignment with the characters in the original image, but other SOTA methods have difficulty aligning images that suffer from distortion caused by overhead shots, which causes upper body dilation and lower body shrinkage. The number under each image represents predicted/ground-truth $f$, FoV angle, and $T_z$.

Table 4: Ablation study of Zolly-H48 of different training settings on 3DPW dataset. w/ PD means whether trained on perspective-distorted datasets (PDHuman, HuMMan), w/ 3DPW means whether fine-tuned on 3DPW [45] dataset. w/ gt $f$ means using ground-truth focal length when fine-tuned on 3DPW.

<table>
<thead>
<tr>
<th>w/ PD</th>
<th>w/ 3DPW</th>
<th>w/ gt $f$</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>PA-MPJPE</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>-</td>
<td>48.3</td>
</tr>
<tr>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>41.3</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>-</td>
<td>40.9</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>47.9</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>40.9</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>39.8</td>
</tr>
</tbody>
</table>

Table 5: Ablation study of Zolly-H48 structure on PDHuman ($\tau = 3.0$). $L_w$ indicates weak-perspective re-projection loss. $L_p$ indicates perspective re-projection loss. $\sum d_j L_w^d$ terms dividing the per-joint distortion weight in our weak-perspective loss.

Table 5: Ablation study of Zolly-H48 structure on PDHuman ($\tau = 3.0$). $L_w$ indicates weak-perspective re-projection loss. $L_p$ indicates perspective re-projection loss. $\sum d_j L_w^d$ terms dividing the per-joint distortion weight in our weak-perspective loss.

5. Conclusion

We present Zolly, the first 3DHMR method that focuses on human reconstruction from perspective distorted images. Our proposed camera model and focal length solution accurately reconstruct the human body, especially in close-view photographs. We introduce a new dataset, PDHuman, and extend two datasets containing perspective-distorted images. Our results show significant value for human mesh reconstruction in perspective-distorted images and can empower many downstream tasks, such as monocular clothed human reconstruction [43, 37, 46] and human motion reconstruction in live shows, vlogs, and selfie videos. Further improvements and broader applications could be explored in the future.

Societal Impacts. Misuse by gaming or animation companies for motion capture can lead to copyright infringement and discourage original content. Fair use advocacy and negotiation with creators can promote sustainable creativity.

Acknowledgement Taku Komura and Wenjia Wang are partly supported by Technology Commission (Ref:ITS/319/21FP) and Research Grant Council (Ref: 17210222), Hong Kong.
References


[34] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In Int. Conf. 3D Vis., 2017. 6, 7


