SPEC: Seeing People in the Wild with an Estimated Camera

Muhammed Kocabas\textsuperscript{1,2}, Chun-Hao Paul Huang\textsuperscript{1}, Joachim Tesch\textsuperscript{1}, Lea Müller\textsuperscript{1}, Otmar Hilliges\textsuperscript{2}, Michael J. Black\textsuperscript{1}

\textsuperscript{1}MPI for Intelligent Systems, Tübingen, \textsuperscript{2}ETH Zurich

Motivation

To reconstruct human bodies given a single image, existing methods assume
1. Intrinsic: weak perspective projection - large focal length $f=5000$
2. Extrinsic: no camera rotation w.r.t. the world

These assumptions lead to below errors in real life images

- CamCalib takes the whole input image as input and predicts camera parameters. Horizon line (green) shows the predicted camera rotation.
- SPEC takes a cropped bounding box as input and extracts image features using a CNN backbone. Predicted camera parameters from CamCalib are concatenated with image features to estimate SMPL body parameters.
- Camera parameters are also taken into account when computing a loss between the projected 3D joints and ground truth.

Method

\begin{itemize}
  \item CamCalib takes the whole input image as input and predicts camera parameters. Horizon line (green) shows the predicted camera rotation.
  \item SPEC takes a cropped bounding box as input and extracts image features using a CNN backbone. Predicted camera parameters from CamCalib are concatenated with image features to estimate SMPL body parameters.
  \item Camera parameters are also taken into account when computing a loss between the projected 3D joints and ground truth.
\end{itemize}

Camera Geometry

To solve this problem, we estimate the camera parameters denoted in (b) from a single image.

\[ R^c = 1, R^d = R^c \neq 1 \]
\[ t^c \neq 0, t^d = 0 \]

Datasets

\begin{itemize}
  \item Pano360 - to train CamCalib
  \item SPEC-MTP - Evaluation only - Subjects are captured from multiple views while mimicking a reference pose. Camera parameters are obtained w.r.t. the global orientation presented in reference poses.
  \item SPEC-SYN - Training and evaluation - Synthetic images AGORA - cameras are randomly sampled
\end{itemize}

Evaluation Metrics

- MPJPE (mean per joint error), PA-MPJPE are the commonly used metric.
- These metrics exists specifically because current methods reconstruct bodies in camera coordinates.
- Instead, we propose variants of MPJPE and PVE that compute the error in world coordinates without the need of camera information and dub them W-MPJPE and W-PVE.

Results on SPEC-MTP

<table>
<thead>
<tr>
<th>Method</th>
<th>SPEC-MTP</th>
<th>SPEC-SYN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPJPE</td>
<td>25.2</td>
<td>34.3</td>
</tr>
<tr>
<td>PA-MPJPE</td>
<td>23.4</td>
<td>32.5</td>
</tr>
<tr>
<td>W-MPJPE</td>
<td>25.2</td>
<td>34.3</td>
</tr>
<tr>
<td>W-PVE</td>
<td>23.4</td>
<td>32.5</td>
</tr>
</tbody>
</table>

Results on SPEC-SYN

<table>
<thead>
<tr>
<th>Method</th>
<th>SPEC-SYN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPJPE</td>
<td>25.2</td>
</tr>
<tr>
<td>PA-MPJPE</td>
<td>23.4</td>
</tr>
<tr>
<td>W-MPJPE</td>
<td>25.2</td>
</tr>
<tr>
<td>W-PVE</td>
<td>23.4</td>
</tr>
</tbody>
</table>

Error Breakdown w.r.t. the Camera Parameters

\begin{itemize}
  \item W-MPJPE breakdown per camera pitch $\phi$:
  \item W-MPJPE breakdown per focal length $f$:
\end{itemize}

Qualitative Results

\begin{itemize}
  \item SPEC-MTP
  \item SPEC-SYN
\end{itemize}