We present the supplementary materials accompanying our manuscript. In Section 1, we compare the convergence speed of the 2D keypoint reprojection loss and the motion loss corresponding to the third paragraph of the introduction. In Section 2, we analyze the training stability in terms of critical hyper-parameters corresponding to the training details of the manuscript. In Section 3, we conduct more analyses on 3DPW [16], including the explanation about two evaluation protocols mentioned in Section 4.1 of the manuscript. In Section 4, we exhibit qualitative results of the BOA on 3DHP [10]. In Section 5, we introduce the details of learning the base model.

1. Convergence Speed Analysis

Corresponding to the description in the third paragraph of the introduction, Figure 1 exhibits the convergence speed of the 2D keypoint reprojection loss \( L_J \) (the first term in Equation (2) of the manuscript) and the motion loss \( L_m \) when jointly learning multiple objectives. Specifically, we randomly choose 100 images, and optimize on each image for 100 times. The mean values of \( L_J \) and \( L_m \) are plotted in Figure 1. We select multiple loss weights (0.1, 0.5, and 1.0) of \( L_m \) for a convincing analysis. The results in Figure 1 demonstrate that \( L_J \) converges faster than \( L_m \). Therefore, in a small number of inference-stage optimization steps\(^1\), the model may learn to fit the pose priors very quickly but then get stuck trying to learn temporal consistency.

2. Training Stability

As mentioned in the Training details (in Section 4 of the manuscript), here we analyze the training stability of the proposed BOA framework. All hyper-parameters of the BOA are listed in Table 1. We analyze the training stability of the BOA in terms of critical hyper-parameters: ones related to temporal constraints \( \mathcal{L}_m \) and \( \mathcal{L}_{mt} \) and the learning rate \( \alpha \) of lower-level weight probe. From Table 2, we found that our model has similar performance in various hyper-parameter settings. Other hyper-parameters (i.e., \( \eta, \beta_1, \gamma_1, \gamma_2, \gamma_3 \)) are finetuned based on the setting of SPIN [8].

3. More Analyses on 3DPW

Further explanation about # PS and # PH. Recall that we mentioned the difference of protocols used in SPIN and HMMR [5] in Section 4.1 of the manuscript. Here we further explain the details. The evaluation results vary significantly in two different protocols. In practice, under the same experimental setting, the PA-MPJPE is 58.8mm using the protocol # PH, while the PA-MPJPE is 49.52mm with

\[^1\text{A common practice [14] is to perform one-step optimization in online adaptation for the sake of inference efficiency.}\]
Table 2: Training stability analysis on critical hyper-parameters: (a) the learning rate $\alpha$ of lower-level update, (b) the loss weight $\mu_1$ of $L_m$, (c) the loss weight $\mu_2$ of $L_{mt}$, (d) the smoothing coefficient $\varepsilon$ of the exponential moving average on the teacher model $T_\omega$ [13], and (e) the interval $\tau$ from the previous image for the motion loss $L_m$. We report both the MPJPE and PA-MPJPE on 3DPW [16] using the protocol of # PS. Note that the optimal parameters are in bold font.

<table>
<thead>
<tr>
<th>Method</th>
<th>Protocol</th>
<th>PA-MPJPE</th>
<th>MPJPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPIN* [4]</td>
<td># PS</td>
<td>146.6</td>
<td></td>
</tr>
<tr>
<td>SMPLify [1]</td>
<td># PS</td>
<td>106.1</td>
<td></td>
</tr>
<tr>
<td>PoseNet3D [15]</td>
<td># PS</td>
<td>63.2</td>
<td></td>
</tr>
<tr>
<td>Song et al. [12]</td>
<td># PS</td>
<td>55.9</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td># PS</td>
<td>49.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Quantitative comparison with skeleton-based models [1, 15, 12] on 3DPW in terms of PA-MPJPE (# PS). Note that the baseline model SPIN* is trained only on Human3.6M.

<table>
<thead>
<tr>
<th>Metric</th>
<th>SPIN</th>
<th>VIBE</th>
<th>BOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVE (mm)</td>
<td>116.4</td>
<td>113.4</td>
<td>91.2</td>
</tr>
</tbody>
</table>

Table 4: Shape evaluation on 3DPW in terms of PVE (mm). Note that the baseline model SPIN* is trained only on Human3.6M.

Comparison with skeleton-based models. We compare with models [1, 15, 12] that taking 2D keypoints as input, and report the PA-MPJPE (# PS) in Table 3. Compared with images, 2D keypoints carry more specific and explicit pose information. As a result, taking 2D keypoint as input is advantageous to pose-related metrics MPJPE and PA-MPJPE [2, 11]. However, relative to the baseline model SPIN* which is also only trained on Human3.6M, the BOA brought more significant improvement than skeleton-based models. Moreover, they use more training data than SPIN* and ours. Even compared with the best skeleton-based model (the fourth row), our model still outperforms 6.4mm in terms of PA-MPJPE. This verifies the advantages of our proposed online adaptation scheme, BOA. Besides, the proposed BOA is also compatible with skeleton-based models.

4. Visualization on 3DHP

In Figure 2, we show qualitative results of the BOA on 3DHP. Even 3DHP has significant differences from Human3.6M [3] in many aspects (e.g., camera parameters, bone length), our BOA still performs well in both wild and indoor scenarios. This verifies the advantages of our bilevel online adaptation framework, which can learn out-of-domain data well.
5. Learning the Base Model on Human3.6M

We further introduce how to learn a base model \( M_{\phi_0} \) on \( D_S \). Although \( D_T \) far sway from training set \( D_S \), some common characteristics is shared, such as body topological structure, kinematic prior. However, taking images as input, \( M_{\phi_0} \) is prone to over-fit on textures. To prevent the learning drift conception, we train the base model in a fully supervised manner. Given an image \( y \in D_S \), the base model \( M \) provides the regression results, including the SMPL parameters \( \{ \hat{\beta}, \hat{\theta} \} \) and the camera parameters \( \Pi \) in a forward pass. According to the pre-defined mesh-to-skeleton mapping in SMPL, we can obtain the estimated 3D keypoints \( \hat{J} \) and its 2D projection \( \hat{j} = \Pi (\hat{J}) \). Then we supervised \( M_{\phi_0} \) as follows:

\[
L_S = \lambda_1 L_J + \lambda_2 L_j + \lambda_3 L_\theta, \tag{1}
\]
\[
L_J = ||J_y - \hat{J}_y||^2, \tag{2}
\]
\[
L_j = ||j_y - \hat{j}_y||^2, \tag{3}
\]
\[
L_\theta = ||\beta - \hat{\beta}||^2 + \lambda_4 ||\theta - \hat{\theta}||^2 \tag{4}
\]

where \( \{ \lambda_1, \lambda_2, \lambda_3, \lambda_4 \} \) are loss weights. The model \( M_{\phi_0} \) follows the architecture of SPIN et al. \[8\] and has the same training setting with them. The only difference is that we exclude the optimization module from training. We follow the same training setting with SPIN. By taking in strong paired 3D supervisions, the base model \( M_{\phi_0} \) can get lots of helpful basic knowledge, e.g. judging body orientation from images, and 2D-to-3D lifting. This point makes the base model possible to be quickly adapted to unseen images.

References

person linear model. *ACM Transactions on Graphics (TOG)*, pages 1–16, 2015. 2


