Bilevel Online Adaptation for Out-of-Domain Human Mesh Reconstruction **Supplementary Material**

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 analysises 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 $\mathcal{L}_J$ (the first term in Equation (2) of the manuscript) and the motion loss $\mathcal{L}_{\boldsymbol{m}}$ when jointly learning multiple objectives. Specifically, we randomly choose 100 images, and optimize on each images for 100 times. The mean values of $\mathcal{L}_J$ and $\mathcal{L}_{\boldsymbol{m}}$ are plot in Figure 1. We select multiple loss weights (0.1,0.5 and 1.0) of $\mathcal{L}_{\boldsymbol{m}}$ for for convincing analysis. The results in Figure 1 demonstrate that $\mathcal{L}_J$ converges faster than $\mathcal{L}_{\boldsymbol{m}}$. 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 $\mathcal{L}_{\boldsymbol{m}}$ and $\mathcal{L}_{\boldsymbol{mt}}$ and the learning rate α of lower-level weight probe. From Table 2, we found that our model has similar performance in various hyper-parameter settings. Other hyper-parameters

[1]A common practice [14] is to perform one-step optimization in online adaptation for the sake of inference efficiency.

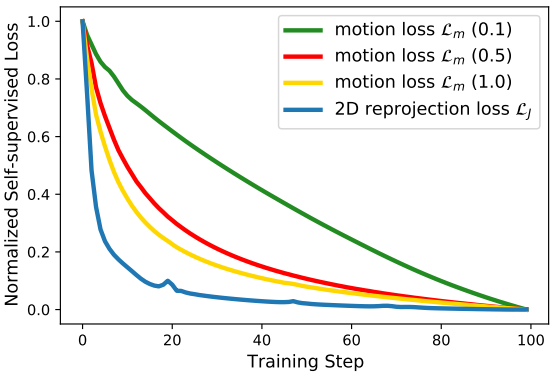

Figure 1: Comparison of convergence speed between $\mathcal{L}_J$ and $\mathcal{L}_{\boldsymbol{m}}$. The numbers in brackets represent loss weights. For a fair comparison, the scale loss is normalized to [0,1].

Notation	Value	Description
μ_1	0.1	The loss weight of motion loss $\mathcal{L}_{\boldsymbol{m}}$
μ_2	0.1	The loss weight of consistent loss $\mathcal{L}_{\boldsymbol{mt}}$
τ	5	The interval from the previous image for $\mathcal{L}_{\boldsymbol{m}}$
δ	0.1	The coefficient of exponential moving average [13]
α	3e-8	The learning rate of lower-level weight probe
η	3e-6	The learning rate of upper-level update
β_1	0.5	The decay rate of first moment in Adam [6]
$\gamma_1, \gamma_2, \gamma_3$	10,1e-4,0.1	The loss weight of the three terms in $\mathcal{L}_F$

Table 1: Hyper-parameters of the BOA. Note that we list the specific values on 3DPW.

(*i.e.* $\eta, \beta_1, \gamma_1, \gamma_2, \gamma_3$) are finetuned based on the setting of SPIN [8].

3. More Analysises 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

	4e-6	6e-6	**8e-6**	1e-5
PA-MPJPE	49.85	51.59	**49.52**	49.91
MPJPE	77.45	81.12	**77.26**	77.79

(a) The learning rate α of lower-level update.

	0.02	**0.1**	0.2	0.4
PA-MPJPE	51.3	**49.5**	51.7	50.7
MPJPE	81.4	**77.3**	82.0	78.9

(b) The loss weight μ_1 of $\mathcal{L}_{\boldsymbol{m}}$.

	0.02	**0.1**	0.2	0.4
PA-MPJPE	49.8	**49.5**	49.6	51.3
MPJPE	77.5	**77.3**	77.4	80.1

(c) The loss weight μ_2 of $\mathcal{L}_{\boldsymbol{mt}}$.

	0.1	0.5	0.7	0.9
PA-MPJPE	**49.5**	50.0	50.4	49.8
MPJPE	**77.3**	78.1	78.4	78.0

(d) The smoothing coefficient ε.

	1	**5**	9	12
PA-MPJPE	51.2	**49.5**	50.8	49.7
MPJPE	80.7	**77.3**	78.9	77.7

(e) The interval τ.

Table 2: Training stability analysis on critical hyper-parameters: (a) the learning rate α of lower-level weight probe, (b) the loss weight μ_1 of $\mathcal{L}_{\boldsymbol{m}}$, (c) the loss weight μ_2 of $\mathcal{L}_{\boldsymbol{mt}}$, (d) the smoothing coeffient δ of the exponential moving average on the teacher model $\mathcal{T}_\omega$ [13], and (e) the interval τ from the previous image for the motion loss $\mathcal{L}_{\boldsymbol{m}}$. 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.

Method	Protocol	PA-MPJPE
SPIN* [4]	# PS	146.6
SMPLify [1]	# PS	106.1
PoseNet3D [15]	# PS	63.2
Song *et al.* [12]	# PS	55.9
Ours	# PS	**49.5**

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.

the protocol # PS. The reason for the significant difference lies in two aspects:

- Choice of SMPL [9] annotation at test time. SPIN uses the SMPL annotation from the official 3DPW as ground-truth. Instead, HMMR adopts the fitted SMPL annotations as ground-truth. Specifically, HMMR uses a neutral template to fit the official SMPL annotations by taking the objective that minimizing vertex error between the fits and the target meshes.
- Choice of the valid test frame. For SPIN, frames in which less than 6 joints are detected are discarded. For HMMR, however, frames need to meet two conditions to be discarded: (1) all detected joints less than 0.1 or the skeleton height less than 0.5 pixels, and (2) all the next frames are also invalid.

Shape evaluation. We adopt the Per Vertex Error (PVE) metric from VIBE [7] to evaluate the shape accuracy of the reconstructed mesh. We take the mesh provided by 3DPW as ground-truth, and the comparison results is shown in Table 4.

Metric	SPIN	VIBE	BOA
PVE (↓)	116.4	113.4	91.2

Table 4: Shape evaluation on 3DPW in terms of PVE (mm).

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.

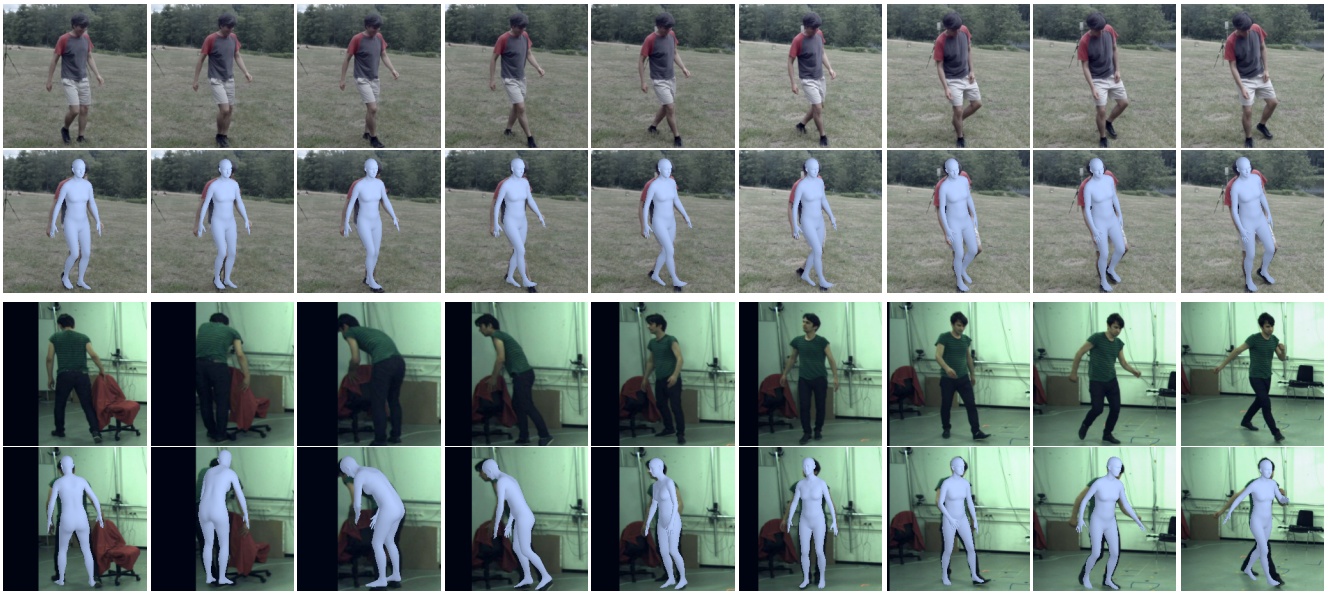

Figure 2: Qualitative results of the BOA on 3DHP. The first and third rows are input sequences. The second and fourth rows show the results.

5. Learning the Base Model on Human3.6M

We further introduce how to learn a base model $\mathcal{M}_{\phi_0}$ on $\mathcal{D}^S$. Although $\mathcal{D}^T$ far sway from training set $\mathcal{D}^S$, some common characteristics is shared, such as body topological structure, kinematic prior. However, taking images as input, $\mathcal{M}_{\phi_0}$ 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 $\boldsymbol{y} \in \mathcal{D}^S$, the base model $\mathcal{M}$ provides the regression results, including the SMPL parameters $\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}\}$ and the camera parameters $\boldsymbol{\Pi}_{\widehat{\psi}}$ in a forward pass. According to the pre-defined mesh-to-skeleton mapping in SMPL, we can obtain the estimated 3D keypoints $\widehat{\boldsymbol{J}}$ and its 2D projection $\widehat{\boldsymbol{j}} = \boldsymbol{\Pi}_{\widehat{\psi}}(\widehat{\boldsymbol{J}})$. Then we supervised $\mathcal{M}_{\phi_0}$ as follows:

$$\mathcal{L}_S = \lambda_1 \mathcal{L}_{\boldsymbol{J}} + \lambda_2 \mathcal{L}_{\boldsymbol{j}} + \lambda_3 \mathcal{L}_{\boldsymbol{\Theta}}, \tag{1}$$

$$\mathcal{L}_{\boldsymbol{J}} = ||\boldsymbol{J}_{\boldsymbol{y}} - \widehat{\boldsymbol{J}}_{\boldsymbol{y}}||_2^2, \tag{2}$$

$$\mathcal{L}_{\boldsymbol{j}} = ||\boldsymbol{j}_{\boldsymbol{y}} - \widehat{\boldsymbol{j}}_{\boldsymbol{y}}||_2^2, \tag{3}$$

$$\mathcal{L}_{\boldsymbol{\Theta}} = ||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}||_2^2 + \lambda_4 ||\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}||_2^2 \tag{4}$$

where $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are loss weights. The model $\mathcal{M}_{\phi_0}$ 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 $\mathcal{M}_{\phi_0}$ 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

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578, 2016. 2

[2] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, pages 769–787, 2020. 2

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, pages 1325–1339, 2014. 2

[4] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar Fine-Tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. 2

[5] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, pages 5614–5623, 2019. 1

[6] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[7] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 2

[8] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 1, 3

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-

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

[10] 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 *3DV*, pages 506–516, 2017. 1

[11] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018. 2

[12] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, pages 744–760, 2020. 2

[13] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 1, 2

[14] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *CVPR*, pages 195–204, 2019. 1

[15] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit Agrawal. Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. In *3DV*, pages 311–321, 2020. 2

[16] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using IMUs and a moving camera. In *ECCV*, pages 601–617, 2018. 1, 2