Bilevel Online Adaptation for Out-of-Domain Human Mesh Reconstruction

Shanyan Guan[*], Jingwei Xu[*], Yunbo Wang[†], Bingbing Ni[†], Xiaokang Yang

Shanghai Jiao Tong University, Shanghai 200240, China

MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China

{shyanguan,xjwxjw,yunbow,nibingbing,xkyang}@sjtu.edu.cn

Abstract

This paper considers a new problem of adapting a pre-trained model of human mesh reconstruction to out-of-domain streaming videos. However, most previous methods based on the parametric SMPL model [36] underperform in new domains with unexpected, domain-specific attributes, such as camera parameters, lengths of bones, backgrounds, and occlusions. Our general idea is to dynamically fine-tune the source model on test video streams with additional temporal constraints, such that it can mitigate the domain gaps without over-fitting the 2D information of individual test frames. A subsequent challenge is how to avoid conflicts between the 2D and temporal constraints. We propose to tackle this problem using a new training algorithm named Bilevel Online Adaptation (BOA), which divides the optimization process of overall multi-objective into two steps of weight probe and weight update in a training iteration. We demonstrate that BOA leads to state-of-the-art results on two human mesh reconstruction benchmarks[1].

1. Introduction

Human mesh reconstruction is a hot topic in computer vision, where improving the generalization ability is one of the major challenges at present. We observe that previous models [21, 38, 25, 26, 1, 59, 15, 43] are prone to overfit the training dataset and usually underperform in out-of-domain testing scenarios. As shown in Figure 1, between different datasets, there usually exist large domain gaps in camera parameters, lengths of body bones, backgrounds, and occlusions, whose negative impact becomes even more severe when we apply the model to streaming data due to the rapidly changing environment of the test domain. In this work, we are interested in finding a good solution to adapting human mesh reconstruction models to out-of-domain
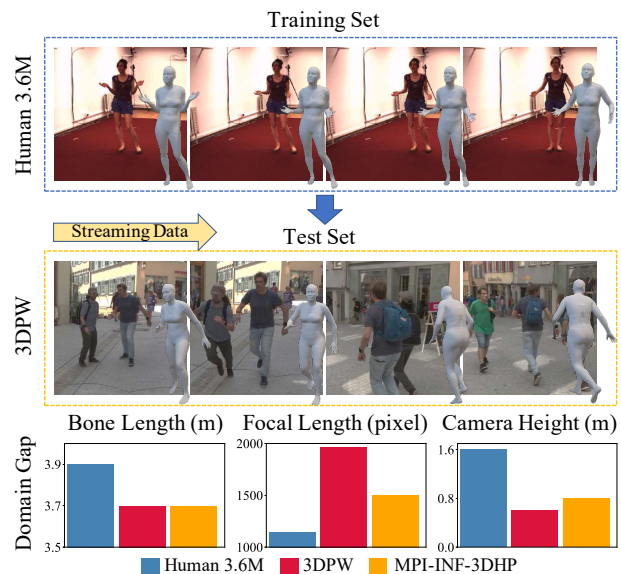

Figure 1. **Top:** Learning to reconstruct human meshes from out-of-domain streaming videos. The main challenges include the domain gaps of plain/crowded backgrounds, with/without occlusions, etc. **Bottom:** Statistics of other typical domain gaps. Here, the bone length refers to the sum of lengths between body joints, whose topology is shared across datasets.

video frames that arrive in a sequential order, which is a practical task in many downstream, real applications, *e.g.*, augmented reality [3], and human-robot interaction [48].

The most serious technical challenge of this task is the lack of 3D annotations of test data. To cope with this problem, some optimization-based approaches [20, 36, 42] learn to update the model on each test frame using frame-wise losses, such as the pose re-projection loss [21, 26] of 2D keypoints[2]. However, the imperfect frame-based loss functions do not always lead to effective online learning directions as expected by the 3D evaluation metric. There is a severe gap between them. As shown in Figure 2, it may cause severe ambiguity in the estimation of depth information,

---

[*]Equal contribution

[†]Corresponding authors: Yunbo Wang, Bingbing Ni

[1]The project website with code and video results is at https://sites.google.com/view/humanmeshboa

[2]It is a common practice to use the ground-truth 2D keypoints for cross-domain human mesh/pose reconstruction
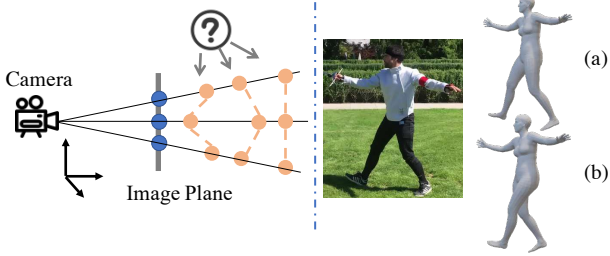
Figure 2. **Left:** Adapting the model with 2D re-projection loss leads to depth ambiguity, where multiple 3D vertices are projected to the same 2D position. **Right:** A showcase that the uncertainty of depth information may lead to the wrong estimation of leg posture.

thus worsening the quality of mesh reconstruction. Moreover, due to the asynchronous arrival of streaming data, the online adaptation model is prone to over-fitting, which will further amplify the difference between the 2D objectives and 3D evaluation metrics.

A straightforward solution is to regularize the training process towards 2D pose objectives using temporal constraints [25, 49, 22], such as the smoothness of mesh reconstruction over time. If the temporal constraints are used properly, the ambiguity of depth estimation can be greatly reduced. However, empirically, a simple combination of 2D losses and temporal constraints tends to obtain undesirable results due to the competition and incompatibility between multiple objectives, in the sense that the gradient of 2D objectives may interfere with the training of the temporal one. Further, solving this problem becomes even more urgent in online adaptation scenarios with streaming data, because without global knowledge of the test domain, the model can easily fall into a sub-optimal solution to either part of the loss functions that is more readily available.

The above two concerns motivate us to tackle the challenging problem of out-of-domain mesh reconstruction from a new perspective. We propose an algorithm named *Bilevel Online Adaptation* (BOA) that greatly benefits joint learning of multiple objectives in this task. It effectively incorporates temporal consistency into the few-step online training by performing bilevel optimization on the streaming test data. Specifically, in BOA, the lower-level optimization step serves as a weight probe to rational model parameters under single-frame pose constraints, while the upper-level optimization step finds a feasible response to overall loss function with temporal constraints. On one hand, our approach avoids overfitting the temporal constraints by retaining the 2D losses for the upper-level optimization. On the other hand, it avoids overfitting the 2D losses by updating the model only at the upper-level optimization step with second-order derivatives. By this means, our approach effectively combines the benefits of pose and temporal constraints. In experiments, we use Human3.6M [19] as the source domain, and take 3DPW [56]

and MPI-INF-3DHP [37] as target domains with streaming video frames. On both benchmarks, our approach consistently outperforms existing approaches [20, 36, 42], showing the excellent ability to tackle notable domain gaps.

2. Problem Setup

A SMPL-based solution to human mesh reconstruction can be usually specified as a tuple of $(\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\Pi}, \mathcal{L})$, where $\boldsymbol{X}$ denotes the observation space[3], and $\boldsymbol{\Theta}$ is the parameter space of SMPL [36]. For each input frame $\boldsymbol{x}_i \in \boldsymbol{X}$, a first-stage model is trained to estimate $\{\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\theta}}_i\} \in \boldsymbol{\Theta}$. Then the SMPL model generates the corresponding mesh and recovers 3D keypoints denoted by $\widehat{\boldsymbol{J}}_i$ using a mesh-to-3D-skeleton mapping pre-defined in SMPL. The third element in the tuple is a weak-perspective projection model for projecting $\widehat{\boldsymbol{J}}_i$ to 2D space, *i.e.*, $\hat{\boldsymbol{j}}_i = \boldsymbol{\Pi}_{\widehat{\psi}_i}(\widehat{\boldsymbol{J}}_i)$, where $\widehat{\psi}_i$ is estimated from $\boldsymbol{x}_i$. The last one in the tuple defines a loss function $\mathcal{L}(\cdot)$ on $(\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\theta}}_i, \widehat{\psi}_i, \widehat{\boldsymbol{J}}_i, \hat{\boldsymbol{j}}_i)$ to learn the first-stage model $\mathcal{M}_\phi$, usually in terms of neural networks.

In this work, we make two special modifications to the above task. First, we focus on out-of-domain scenarios, in the sense that large discrepancies may exist between the data distributions of the source training domain $\mathcal{D}^{\text{tr}}$ and the target test domain $\mathcal{D}^{\text{test}}$. Second, we specifically focus on dealing with streaming video frames at test time. These changes bring in two challenges: (1) The ground truth values of the target domain parameters in $(\boldsymbol{\Theta}, \boldsymbol{\Pi})$ are always unavailable throughout the learning process, which is different from standard online learning. (2) The distribution of the target domain is difficult to be estimated because the frames $\{\boldsymbol{x}_i\}_{i=1}^N \in \boldsymbol{X}$ are sequentially available and their distributions are continuously changing, which is different from standard domain adaptation setups.

3. Bilevel Online Adaptation

In this section, we first formalize the online adaptation framework as a solution to out-of-domain human mesh reconstruction from streaming sequential data. To make the online adaptation more effective, we then propose a bilevel optimization algorithm that incorporates unsupervised temporal constraints into the training paradigm.

3.1. Online Adaptation Framework

Unlike the existing approaches [21, 9, 49] that try to solve the problem introduced in Section 2 by learning more generalizable features in the source domain $\mathcal{D}^{\text{tr}}$, we here present an alternative solution that performs online test-time training directly on the target domain $\mathcal{D}^{\text{test}}$. A potential benefit is that as it is solely performed on $\mathcal{D}^{\text{test}}$, it can be jointly

---

[3]We here consider $\{\boldsymbol{x}_i\}_{i=1}^N \in \boldsymbol{X}$ as a set of consecutive video frames.

Algorithm 1 Bilevel Online Adaptation

Input: sequential frames $\{\boldsymbol{x}_i\}_{i=1}^N$ from test set, the base model $\mathcal{M}_{\phi_0}$ with model parameter ϕ_0, a teacher model $\mathcal{T}_{\omega_0}$ with model parameter ω_0, learning rates α and η.

Output: SMPL parameters $\{\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\theta}}_i\}_{i=1}^N$, camera parameters $\{\widehat{\boldsymbol{\psi}}_i\}_{i=1}^N$

1: **Initialize** $\omega_0 \leftarrow \phi_0$
2: **for** $i = 1, \ldots, N$ **do**
3: $\quad \phi_{i-1}^0 \leftarrow \phi_{i-1}$
4: $\quad$ **for** $t = 1, \ldots, T$ **do** $\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ ▷ For each bilevel optimization step
5: $\quad\quad \mathcal{L}_{\text{low}} \leftarrow \mathcal{L}_F(\boldsymbol{x}_i; \phi_{i-1}^{t-1})$ $\qquad\qquad\qquad$ ▷ Calculate lower-level loss with single-frame pose constraints
6: $\quad\quad \phi_{i-1}^{(t-1)\prime} \leftarrow \phi_{i-1}^{t-1} - \alpha \nabla_\phi \mathcal{L}_{\text{low}}$ $\qquad\qquad$ ▷ Probe for rational weights under pose constraints
7: $\quad\quad \mathcal{L}_{\text{up}} \leftarrow \mathcal{L}_F(\boldsymbol{x}_i; \phi_{i-1}^{(t-1)\prime}) + \mathcal{L}_T(\boldsymbol{x}_i; \omega_{i-1}, \phi_{i-1}^{(t-1)\prime})$ $\qquad$ ▷ Calculate upper-level loss with temporal constraints
8: $\quad\quad \phi_{i-1}^t \leftarrow \phi_{i-1}^{t-1} - \eta \nabla_\phi \mathcal{L}_{\text{up}}$ $\qquad\qquad\qquad$ ▷ Update model weights with second-order derivatives
9: $\quad$ **end for**
10: $\quad \phi_i \leftarrow \phi_{i-1}^T$
11: $\quad \omega_i \leftarrow \delta\omega_{i-1} + (1-\delta)\phi_i$ $\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ ▷ Update the teacher model
12: $\quad \{\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\theta}}_i\}, \boldsymbol{\Pi}_{\widehat{\boldsymbol{\psi}}_i} \leftarrow \mathcal{M}_{\phi_i}(\boldsymbol{x}_i)$ $\qquad$ ▷ Estimate the SMPL parameters and the camera parameter of $\boldsymbol{x}_i$ using $\mathcal{M}_{\phi_i}$
13: **end for**

---

used with the state-of-the-art approaches that learn generalizable features from $\mathcal{D}^{\text{tr}}$ to further improve the quality of transfer learning.

Alg. 1 shows the proposed online adaptation framework. Here we denote the pre-trained model from the source domain as $\mathcal{M}_{\phi_0}$. Our framework does not have special requirements for the pre-training method, but typically, $\mathcal{M}_{\phi_0}$ is trained offline to regress the ground truth SMPL parameters in a fully supervised manner. Given sequentially arrived target video frames $\{\boldsymbol{x}_i\}_{i=1}^N \in \mathcal{D}^{\text{test}}$, a straight forward solution to quickly absorbing the domain-specific knowledge is to fine-tuning $\mathcal{M}$ continuously on each individual $\boldsymbol{x}_i$, following the online adaptation paradigm proposed by Tonioni *et al.* [52]. We take it as a baseline algorithm that computes the unsupervised loss function $\mathcal{L}$ with pose constraints[4] on each $\boldsymbol{x}_i$, and performs a single optimization step as follows before the inference step:

$$\phi_i = \phi_{i-1} - \alpha \nabla_\phi \mathcal{L}(\boldsymbol{x}_i; \mathcal{M}_{\phi_i}), \qquad (1)$$

where α is the learning rate of gradient descent.

A potential disadvantage of the baseline algorithm is that although fine-tuning a learned model on unlabeled target data may help to handle rapidly changing test environments, an imperfect unsupervised loss function may lead to wrong directions of the one-step gradient descent, and may harm the overall algorithm. It may cause catastrophic overfitting to some undesirable information of current observation that is unrelated to the reconstruction quality. To alleviate this issue, we propose the following spatiotemporal bilevel optimization approach.

---

[4]We discuss more about the specific forms of $\mathcal{L}$ in Section 4.

3.2. Spatiotemporal Bilevel Optimization

Considering the setting of out-of-domain streaming data, *i.e.*, video frames arrive at a sequential order, there generally exists strong temporal dependency between frames, which can be leveraged to improve the quality of online adaptation. Let us suppose that we have two objectives respectively for frame-wise constraints and temporal consistency, denoted by $\mathcal{L}_F$ and $\mathcal{L}_T$, whose specific forms will be discussed later. Straightforward approaches to combine $\mathcal{L}_F$ and $\mathcal{L}_T$ include jointly optimizing them by adding them together or iteratively performing two-stage optimization (Figure 2b). However, these methods usually lead to suboptimal results due to the competition and incompatibility between the objectives, in the sense that the gradient of the single-frame constraint may interfere with the training of the temporal one. We also observe that the single-frame constraint is usually optimized much faster than the temporal one. That is to say, in a small number of inference-stage optimization steps, the model may learn pose priors very quickly but then get stuck trying to learn temporal consistency. Therefore, the first and foremost challenge we confront is to design an online optimization scheme to prevent overfitting to each objective and maximize the power of both single-frame and temporal constraints.

Lower-level weight probe with single-frame constraints. We formulate the problem of identifying effective model weights under spatiotemporal multi-objectives as a bilevel optimization problem. In this setup, as shown in Figure 3, the lower-level optimization step serves as a weight probe to rational models under single-frame pose constraints, while the upper-level optimization step finds a feasible response to temporal constraints. Specifically, for the i-th test sam-
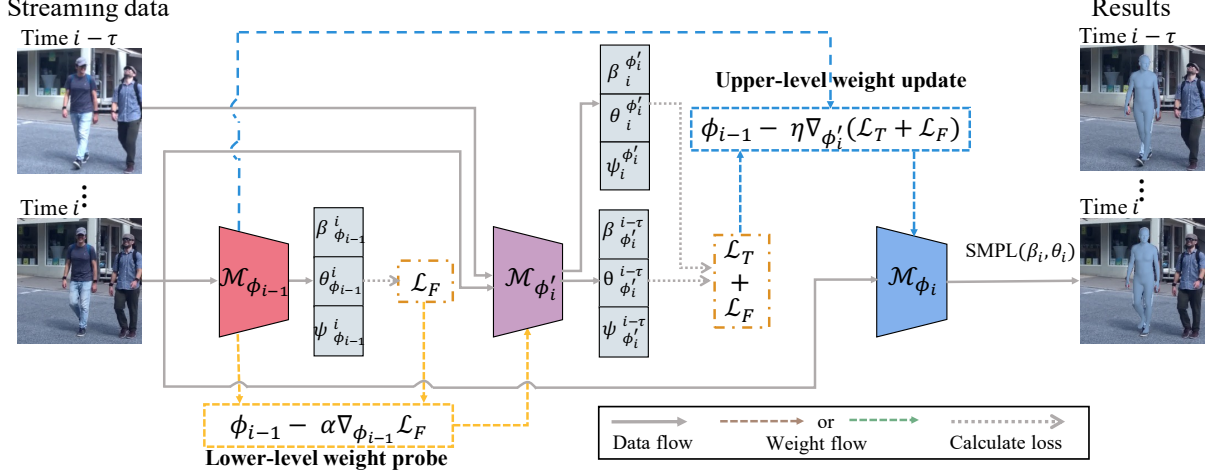
Figure 3. A diagram of the proposed bilevel online adaptation method. For simplicity, we only show one iteration of bilevel optimization. The lower-level training step serves as a parameter *probe* to find a feasible response to the frame-wise pose constraints. The upper-level training step minimizes the overall multi-objectives in space-time and updates the model with second-order derivatives.

ple, the model from the last online adaptation step, denoted by $\mathcal{M}_{\phi_{i-1}}$, is firstly optimized with the single-frame constraints, $\mathcal{L}_F$, to obtain a set of temporary weights denoted by ϕ_i'. We name this procedure as the *lower-level probe* (Lines 5-6 in Alg. 1), in the sense that first, ϕ_i' can be feasible responses to the easy component of multi-objectives, which best facilitates the rest of the learning procedure for temporal consistency; Second, ϕ_i' is not directly used to update $\mathcal{M}_{\phi_{i-1}}$. At this level we focus on the spatial constraints on individual frames:

$$\mathcal{L}_F = \gamma_1 ||\boldsymbol{j}_i - \widehat{\boldsymbol{j}}_i||_2^2 + \gamma_2 \rho(\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\theta}}_i) + \gamma_3 \mathcal{L}_S, \quad (2)$$

where $\{\gamma_1, \gamma_2, \gamma_3\}$ are the loss weights. The first term in $\mathcal{L}_F$ is a straightforward supervision of the re-projection error of 2D keypoints. The second term is the prior constraint on the shape and pose parameters, which is a common practice in mesh reconstruction. $\rho(\cdot)$ calculates the distance of the estimated $\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\theta}}_i$ to their statistic priors[5]. The third term is the fully supervised loss with 3D keypoints on a randomly sampled source data, which has two benefits: (1) preventing the catastrophic forgetting of the basic knowledge learned from $\mathcal{D}^{tr}$. (2) providing the online updated model a continuous 3D supervision to keep it from overfitting the imperfect unsupervised loss functions. After optimization at the lower level, we obtain the probe model $\mathcal{M}_{\phi_i'}$ for subsequent upper-level learning. Note that, due to a lack of 3D supervisions in the target domain, the above $\mathcal{L}_F$ is insufficient to recover the 3D body. Therefore, it is essential to explore temporal correlations in streaming data to reduce the ambiguity of mesh construction.

---
[5]These priors are obtained from a commonly-used third-party database.

Upper-level weight update with temporal constraints.
At an upper optimization level, we calculate the overall spatiotemporal multi-objectives using $\mathcal{M}_{\phi_i'}$ obtained at the lower-level optimization step, and then back-propagate with second-order derivatives to update the original ϕ_{i-1}, as shown in Lines 7-8 in Alg. 1. As for the specific form the motion constraints, given two images $\boldsymbol{x}_i$, $\boldsymbol{x}_{i-\tau}$ at an interval τ with their 2D keypoints $\boldsymbol{j}_i, \boldsymbol{j}_{i-\tau}$ and the estimated $\widehat{\boldsymbol{j}}_i$, $\widehat{\boldsymbol{j}}_{i-\tau}$, the motion loss is defined as

$$\mathcal{L}_m = ||\widehat{\boldsymbol{m}}_i - \boldsymbol{m}_i||_2^2, \quad (3)$$

where $\boldsymbol{m}_i = \boldsymbol{j}_i - \boldsymbol{j}_{i-\tau}$, and $\widehat{\boldsymbol{m}}_i = \widehat{\boldsymbol{j}}_i - \widehat{\boldsymbol{j}}_{i-\tau}$. Note that both $\widehat{\boldsymbol{j}}_i$ and $\widehat{\boldsymbol{j}}_{i-\tau}$ are obtained from the probe model $\mathcal{M}_{\phi_i'}$. Furthermore, we maintain an exponential moving average of history models with a teacher model (similar to Mean-Teacher [50]), denoted by $\mathcal{T}_\omega$. We regularize the output of $\mathcal{M}_{\phi_i'}$ to be consistent with $\mathcal{T}_{\omega_{i-1}}$:

$$\mathcal{L}_{mt} = ||\mathcal{T}_{\omega_{i-1}}(\boldsymbol{x}) - \mathcal{M}_{\phi_i'}(\boldsymbol{x})||_2^2, \quad (4)$$

which is then combined with the motion loss to obtain the overall temporal constraints that focus on the consistency of both the sequentially updated model weights and the reconstruction results as well:

$$\mathcal{L}_T = \mu_1 \mathcal{L}_m + \mu_2 \mathcal{L}_{mt}, \quad (5)$$

where μ_1 and μ_2 control the weights of the two temporal loss terms. From another perspective, these two losses are complementary with each other: the teacher model $\mathcal{T}$ maintains long-term temporal information, and the motion loss $\mathcal{L}_m$ is a constraint on short-term motion consistency.

Alternatives for online adaptation schemes. As briefly mentioned above, there are several single-level optimization alternatives of the spatiotemporal multi-objectives, *e.g.*, (1) one-stage joint adaptation: online adapting the model with a combined loss of $\mathcal{L}_F + \mathcal{L}_T$. (2) two-stage adaptation: adapting the model iteratively with $\mathcal{L}_F$ and $\mathcal{L}_T$ in a cascaded optimization manner. However, we observe that the joint adaptation scheme is prone to lead to ineffective training of the temporal constraints due to the incompatibility between multiple objectives. The two-stage scheme adapts the model to individual frames under the single-frame constraints repeatedly, which commonly leads to severe overfitting and drifting away from the final 3D reconstruction metric. The key insights of BOA are as follows: First, it avoids overfitting the temporal constraints by retaining the pose prior loss for the upper optimization level. Second, it avoids overfitting the pose priors by updating the model weights only at the upper optimization level with second-order derivatives. By this means, BOA effectively combines the profits of both single-frame and temporal constraints, achieving considerable improvement over its alternatives.

Network architectures. Following the majority of previous SMPL-based human reconstruction models, we use a ResNet-50 [16] pre-trained on ImageNet [8] for encoding individual video frames. The encoded features are then delivered to two fully-connected layers with 1,024 neurons, followed by a dropout layer [47]. The final layer of $\mathcal{M}_\phi$ is a fully-connected layer with 85 neurons. During streaming adaptation, only one image is taken as input. As a result, we replace Batch Normalization [18] with Group Normalization [58] to estimate more accurate statistics.

4. Experiments

Datasets. We use the Human3.6M dataset for training the source model and learn to adapt the model to the 3DHP and 3DPW datasets. Table 1 presents the statistics of typical domain gaps among these datasets.

- **Human3.6M [19]** is captured in a controlled environment, which has 11 subjects in total. Following the previous approaches [25, 21], we train the base model on 5 subjects (S1, S5, S6, S7, S8), and down-sample all videos from 50fps to 10fps.

- **3DHP [37]** is the test split of the MPI-INF-3DHP dataset. It consists of 2,929 valid frames from 6 subjects performing 7 actions, collected from both indoor and outdoor environments.

- **3DPW [56]** is a multi-person dataset captured by a handheld camera, where most videos are collected from outdoor environments. As 3DHP, we also use the test set of 3DPW as a streaming target domain.

Dataset	Focal len. (pixel)	Bone len. (m)	Camera dist. (m)	Camera ht. (m)
H3.6M	1146.8±2.0	3.9±0.1	5.2±0.8	1.6±0.1
3DPW	1962.2±1.5	3.7±0.1	3.5±0.7	0.6±0.8
3DHP	1497.9±2.8	3.7±0.1	3.8±0.8	0.8±0.4

Table 1. Typical domain gaps among datasets in terms of focal length, bone length, camera distance, and camera height [57].

Training details. We first train the base model $\mathcal{M}$ on the Human3.6M dataset and take 3DPW and 3DHP as test sets. All video frames are cropped and then scaled to 224×224 pixels according to the bounding boxes calculated from 2D keypoints. For the base model $\mathcal{M}$, we follow the same training scheme as SPIN (more details can be found in [26]). For the training of BOA on 3DPW, we choose the Adam optimizer [24] with the learning rate $\eta = 3e^{-6}$ ($\beta_1 = 0.5, \beta_2 = 0.999$). The loss weights in $\mathcal{L}_F$ are $\gamma_1 = 10, \gamma_2 = 1$, and $\gamma_3 = 0.1$. The loss weights in $\mathcal{L}_T$ are $\mu_1 = 0.1$ and $\mu_2 = 0.1$. As for 3DHP, the learning rate is set to $2e^{-6}$ ($\beta_1 = 0.2, \beta_2 = 0.999$). We set $\gamma_1 = 10, \gamma_2 = 1, \gamma_3 = 0.1$ in $\mathcal{L}_F$ and $\mu_1 = 0.1, \mu_2 = 0.1$ in $\mathcal{L}_T$. Note that the order of streaming videos in 3DPW and 3DHP is pre-defined (same for all compared models), and the batch size of online optimization is 1. We set $T = 1$ (Alg. 1) for the efficiency of adaptation. Please refer to the supplementary material for more analyses of hyper-parameters.

Baselines. We initially compare BOA with end-to-end methods, including frame-based methods [21, 27, 7, 26, 7, 38], video-based methods [22, 25], and those attempting to learn generalizable features from the training domain, such as Sim2Real [9] and DSD-SATN [49]. Given a video frame, end-to-end methods directly estimate its SMPL parameters. We also include existing approaches that fine-tune SMPL parameters $\boldsymbol{\beta}, \boldsymbol{\theta}$ [5, 2] or model parameters ϕ [20] on the target domain. Different from these approaches, BOA adapts ϕ_i in an online fashion, which is more challenging. Please refer to the supplementary material for more details.

Evaluation metrics. Following previous works [21, 9, 61], we evaluate our model in terms of Mean Per Joint Position Error (MPJPE), Procrustes-Aligned MPJPE (PA-MPJPE), and the Percentage of Correct Keypoints (PCK) with a threshold of 150mm on 3DHP.

4.1. Quantitative Evaluation

Results on 3DPW. Table 2 presents quantitative comparisons on 3DPW in MPJPE and PA-MPJPE. Following HMMR or SPIN, most existing methods adopt two kinds of pre-processing protocols on 3DPW as illustrated in Table 3. These two protocols have significant differences in the number of test images and SMPL annotations, which have a great impact on the evaluation. Please refer to the

Figure 4. A qualitative comparison of mesh reconstruction on 3DPW streaming data. We zoom in on the limbs for better visualization.

Method	Prot.	PA-MPJPE$\downarrow$	MPJPE$\downarrow$
HMR [21]	#PH	76.7	130.0
Sim2Real [9]	#PH	74.7	-
GraphCMR [27]	#PS	70.2	-
SPIN [26]	#PS	59.2	96.9
I2L-MeshNet [38]	#PS	58.6	93.2
Pose2Mesh [7]	#PS	58.9	89.2
DSD-SATN [49]	#PS	69.5	-
HMMR [22]	#PH	73.6	116.5
SMPLify [5]	#PH	106.1	199.2
Arnab *et al.* [2]	#PH	72.2	-
EFT [20]	#PS	55.7	-
BOA	#PH	58.8	92.1
BOA	#PS	**49.5**	**77.2**

Table 2. Results on 3DPW, including end-to-end approaches (top) and those fine-tuned on the target domain (middle).

Prot.	SMPL annotation	#Valid frames
#PS (SPIN)	Original	35,515
#PH (HMMR)	The fits	26,234

Table 3. Different protocols of pre-processing the 3DPW data by SPIN [26] and HMMR [22]. #PS uses the SMPL annotations from the original 3DPW as labels, while #PH uses the fitted neutral results (without gender information) as labels.

supplementary materials for more details. In Table 2, we mark the protocol used in the original literature of each compared method. Compared with other end-to-end methods (top part), BOA achieves better performance in both #PS and #PH, and particularly outperforms the methods that are designed to learn generalizable features at training time [49, 21, 9], which indicates that our test-time adaptation approach can better mitigate the domain gap by properly exploiting the streaming data from the test domain. Besides, we also observe that BOA outperforms the compared models that are fine-tuned on the entire training set of 3DPW in an offline manner (middle part). Note that BOA

Method	3DHP-$\mathcal{D}^{\text{train}}$	PA-MPJPE$\downarrow$	MPJPE$\downarrow$	PCK$\uparrow$
Vnect	✓	98.0	124.7	83.9
HMR	✓	113.2	169.5	77.1
SPIN	✓	80.4	124.8	87.0
BOA	✗	**77.4**	**117.6**	**90.3**

Table 4. Results on 3DHP. All models but BOA are trained on the training split of MPI-INF-3DHP, while BOA performs best.

does not require access to the training set. In addition, we do not include the results from VIBE [25] (56.5 in PA-MPJPE and 93.5 in MPJPE) in quantitative comparison, since it was evaluated on the same number of test images under #PH but uses the same SMPL annotations under #PS.

Results on 3DHP. Table 4 gives the MPJPE and PA-PMJPE results on 3DHP, which is the test set of the entire MPI-INF-3DHP domain. Note that all models but BOA are directly trained on the training set $\mathcal{D}^{\text{train}}$ of MPI-INF-3DHP in an offline fashion, in the sense that the global knowledge from the test domain is more accessible to these compared models. Although BOA has never been trained on $\mathcal{D}^{\text{train}}$, it still performs best on the corresponding test split, showing a strong adaptability to a rapidly changing test environment.

4.2. Qualitative Evaluation

Figure 4 presents a typical showcase of mesh reconstruction on the challenging 3DPW dataset. The first row refers to human meshes generated by VIBE [25], while the second row corresponds to our results. We zoom in on the limbs for better visualization and observe that the reconstruction quality of VIBE is less satisfying, *e.g.*, the positions of arms and legs are not correctly estimated. By contrast, our model can capture the depth structure of the human subject, which is mainly due to the proposed bilevel optimization scheme and spatiotemporal constraints. Figure 5 presents a sequence of input videos with severe occlusions, where the subject of interest (the man in the middle)
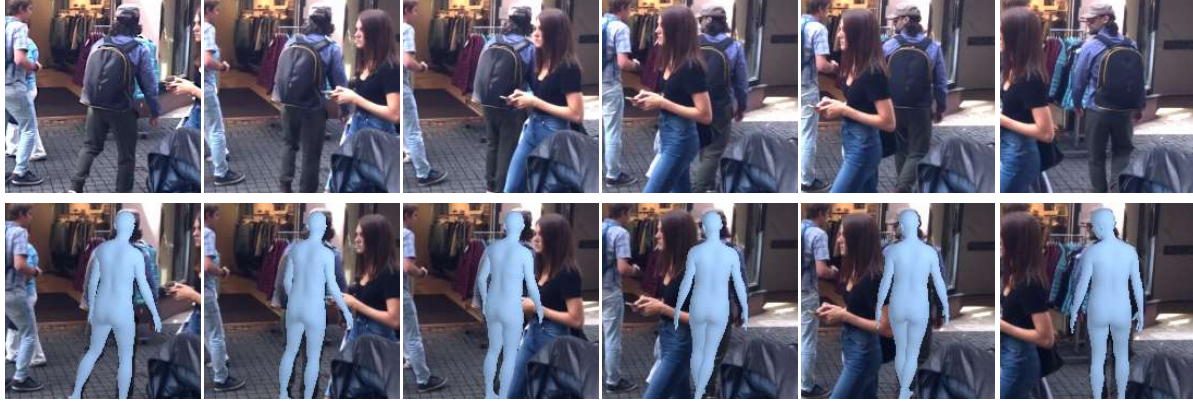
Figure 5. Reconstruction results under severe occlusion. First row: the input frames. Second row: mesh reconstruction results. The subject of interest (the man in the middle) is occluded by the walking woman, which is challenging for out-of-domain mesh reconstruction.
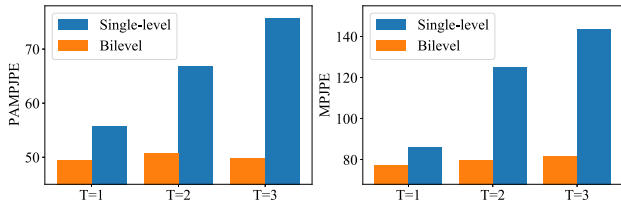

Figure 6. Ablations on the number of optimization steps T (#PS on 3DPW). BOA has a more stable performance with the growth of T, compared with the its single-level counterpart.

is covered by the walking woman. Still, BOA successfully estimates the occluded human body.

4.3. Ablation Study

Ablations on the number of optimization steps. With the growth of the optimization steps (T in Alg. 1), as shown in Figure 6, the error of single-level training scheme, which combines $\mathcal{L}_F$ and $\mathcal{L}_T$ in a multi-objective, in both PA-MPJPE and MPJPE increases quickly. For comparison, the performance of BOA decreases at a much slower rate, which indicates that the single-level optimization is more likely to result in over-fitting to the current video frame, and thus makes it difficult for the model to quickly adapt to the next frame. This effect can be greatly alleviated by the proposed bilevel optimization method.

Ablations on online optimization schemes. As shown in Table 5, we investigate the effectiveness of the proposed bilevel adaptation scheme and compare it with other variants. Specifically, **B1-B3** refer to the single-level, one-stage optimization scheme, while **B4-B5** are trained by updating model parameters with alternate loss functions (*i.e.*, $\mathcal{L}_{low}$ and $\mathcal{L}_{up}$). Note that the major difference between two-stage and bilevel is whether the parameters in $\mathcal{M}_{\phi_i}$ are obtained from $\mathcal{M}_{\phi'_{i-1}}$ or $\mathcal{M}_{\phi_{i-1}}$. We observe that, despite the use of temporal constraints, **B3** performs worse than **B1**, indicating that the straightforward combination of multi-objectives

Index	Optim.	$\mathcal{L}_{low}$	$\mathcal{L}_{up}$	PA-MPJPE	MPJPE
B1	1-stage	$\mathcal{L}_F$	-	55.7	86.0
B2	1-stage	$\mathcal{L}_T$	-	140.1	245.5
B3	1-stage	$\mathcal{L}_F, \mathcal{L}_T$	-	58.9	94.5
B4	2-stage	$\mathcal{L}_F$	$\mathcal{L}_T$	59.3	91.5
B5	2-stage	$\mathcal{L}_F$	$\mathcal{L}_F, \mathcal{L}_T$	55.2	85.1
B6	Bilevel	$\mathcal{L}_F$	$\mathcal{L}_T$	142.5	257.1
Final	Bilevel	$\mathcal{L}_F$	$\mathcal{L}_F, \mathcal{L}_T$	49.5	77.3

Table 5. Ablation studies on the training schemes (#PS on 3DPW). **B1-B3** update model parameters with constant forms of losses in a single optimization step. **B4-B5** update model parameters with alternate loss functions. **B6** and **Final** use bilevel optimization.

Index	$\mathcal{L}_F$	$\mathcal{L}_m$	$\mathcal{L}_{mt}$	PA-MPJPE	MPJPE
B7	✓	✗	✓	53.0	81.8
B8	✓	✓	✗	51.7	82.1
Final	✓	✓	✓	49.5	77.3

Table 6. Ablations on BOA temporal constraints (#PS on 3DPW).

leads to sub-optimal results. Even though we use the multi-objectives in a two-stage training scheme (**B5**), we can only observe a minor improvement over the vanilla **B1** model. By contrast, the final proposed BOA is shown to effectively combine the best of both constraints and achieve considerable improvement over all compared baselines.

Ablations on temporal constraints. Table 6 shows the ablation studies for the proposed two temporal constraints $\mathcal{L}_{mt}$ and $\mathcal{L}_m$. By comparing **B7** with **Final**, we can observe that the ues of $\mathcal{L}_m$ reduces PA-MPJPE from 53.0mm to 49.5mm, while the use of $\mathcal{L}_T$ **B8** reduces PA-MPJPE from 51.7mm to 49.5mm. A possible reason is that the motion loss $\mathcal{L}_m$ focuses on short-term temporal constraint and helps to recalibrate pose artifacts relative to the last frame. By comparing **B8** with **Final**, we can find that the use of $\mathcal{L}_{mt}$ significantly reduces MPJPE, which indicates that the
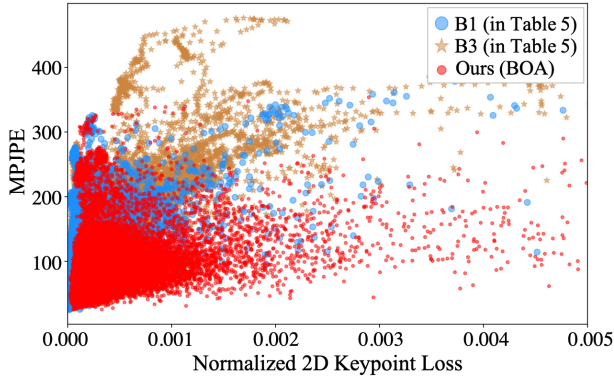
Figure 7. Correlations between the 2D pose re-projection loss and the evaluation metric. **B1** uses frame-based losses only. **B3** optimizes the frame-based and temporal losses by integrating them in a unified loss function. Only with **BOA**, the MPJPE results are consistent with the frame-based loss, showing that BOA greatly reduces the 3D ambiguity of estimated meshes.

long-term information carried by $\mathcal{L}_{mt}$ is beneficial for consistent mesh reconstruction. It may help to mitigate the domain gaps caused by systematic biases such as the focal length and camera orientations.

Analyses on loss-metric correlations. In Figure 7, the X-axis refers to the normalized 2D keypoint loss and the Y-axis is the MPJPE. The blue dots are the results of the vanilla baseline only trained with frame-based loss functions (**B1**), the yellow stars correspond to the baseline model trained with multi-objectives (**B3**), and the red dots indicate bilevel online adaptation. We can see that, first, although a straightforward use of temporal constraints (**B3**) achieves comparable results in the 2D loss with **B1**, it harms the 3D evaluation metric. Second, with BOA, the model achieves MPJPE results that are more consistent with the frame-based loss, indicating that BOA can reduce 3D ambiguity by using the temporal constraints more appropriately.

5. Related Work

3D human mesh reconstruction. SMPL [36] is a widely used parametric model for 3D human mesh reconstruction, which is also adopted in this work. The early methods [14, 46, 5, 28, 17, 60] generally adopt the optimization scheme, where a standard T-Pose SMPL model is gradually fit to an input image according to the silhouettes [28] or 2D keypoints [5]. These optimization-based methods are time-consuming, *i.e.*, they often struggle to reduce the inference time spent on a single input. Recently, many approaches [21, 38, 25, 26, 1, 59, 15, 43] use deep neural networks to regress the parameters of the SMPL model, which are efficient and accurate if large-scale data is available. The major drawback of CNN-based regression models is the generalization ability. For example, deep models trained on

indoor dataset generally do not have satisfying results [21] if tested on an in-the-wild dataset. To tackle this problem, Kanazawa *et al.* [21] propose an adversarial framework, utilizing the unpaired 3D annotations, to facilitate the reconstruction. Several researches [49, 53, 33, 40, 44] also show that the paired 3D annotation is not necessary, attempting to find more representative temporal features [49, 53] or employ more informative input such as RGB-D [33], and part segmentation [40, 44] to facilitate human mesh reconstruction. However, there still exists a principled challenge in this task, where neither the unpaired 3D annotation nor the other mentioned intermediate representations could effectively fill the gap between two largely different datasets. In this work, we propose to tackle this problem by using an online adaptation algorithm, named BOA. The key insight is BOA exploits the time constraints of test frames while avoiding overfitting with bilevel optimization.

Unsupervised online adaptation. Unsupervised online adaptation refers to sequentially adapting a pre-trained model at test time in an unsupervised manner. It is an emerging technique to prevent model crashing when the test data is diverse from the training data. Previous methods [10, 6, 4, 35, 41, 54, 52, 6, 51, 62, 34] use it for tasks other than mesh reconstruction, such as video segmentation [54], tracking [41], and stereo matching [52, 51]. In this paper, we present a pilot study of unsupervised online adaptation in the context of human mesh reconstruction. Beyond unsupervised online adaptation, many previous approaches effectively learn generalizable features through meta-learning [12, 11], extracting domain-invariant representations [23, 39, 13, 29, 32, 30, 31], or learning with adversarial examples [45, 55] without requiring access to target labels. However, none of these approaches focus on how to adapt a pre-trained model to streaming data.

6. Conclusion

In this paper, we presented a new research problem of reconstructing human meshes from out-of-domain streaming videos. We proposed a new online adaption algorithm that learns temporal consistency with bilevel optimization and demonstrated that it can greatly benefit the multi-objective training process in space-time. Our approach outperforms the state-of-the-art mesh reconstruction methods on two benchmarks with rapidly changing test environments.

Acknowledgement

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3D human motion modelling. In *ICCV*, pages 7144–7153, 2019. 1, 8

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, pages 3395–3404, 2019. 5, 6

[3] Mark Billinghurst. Introduction to Augmented Reality. In *VR*, page 266, 2004. 1

[4] Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. In *ICLR*, 2018. 8

[5] 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. 5, 6, 8

[6] Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational bayes. In *NeurIPS*, pages 1727–1735, 2013. 8

[7] 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. 5, 6

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[9] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In *NeurIPS*, pages 12949–12961, 2019. 2, 5, 6

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 2011. 8

[11] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *AISTATS*, pages 1082–1092, 2020. 8

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 8

[13] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoderss. In *ICCV*, pages 2551–2559, 2015. 8

[14] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*, pages 1381–1388, 2009. 8

[15] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, pages 10884–10894, 2019. 1, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[17] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, pages 421–430, 2017. 8

[18] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 5

[19] 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, 5

[20] 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. 1, 2, 5, 6

[21] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 2, 5, 6, 8

[22] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, pages 5614–5623, 2019. 2, 5, 6

[23] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*, pages 158–171, 2012. 8

[24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[25] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 1, 2, 5, 6, 8

[26] 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, 5, 6, 8

[27] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 5, 6

[28] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, pages 6050–6059, 2017. 8

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017. 8

[30] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 8

[31] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, pages 1446–1455, 2019. 8

[32] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018. 8

[33] Ren Li, Changjiang Cai, Georgios Georgakis, Srikrishna Karanam, Terrence Chen, and Ziyan Wu. Towards robust RGB-D human mesh recovery. *arXiv preprint arXiv:1911.07383*, 2019. 8

[34] Shunkai Li, Xin Wang, Yingdian Cao, Fei Xue, Zike Yan, and Hongbin Zha. Self-supervised deep visual odometry

with online adaptation. In *CVPR*, pages 6339–6348, 2020. 8

[35] Hong Liu, Mingsheng Long, Jianmin Wang, and Yu Wang. Learning to Adapt to Evolving Domains. In *NeurIPS*, 2020. 8

[36] 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. 1, 2, 8

[37] 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. 2, 5

[38] Gyeongsik Moon and Kyoung Mu Lee. I2l-MeshNet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 1, 5, 6, 8

[39] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18, 2013. 8

[40] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, pages 484–494, 2018. 8

[41] Eunbyung Park and Alexander C Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In *ECCV*, pages 569–585, 2018. 8

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*, pages 10975–10985, 2019. 1, 2

[43] 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. 1, 8

[44] Nadine Rueegg, Christoph Lassner, Michael J Black, and Konrad Schindler. Chained representation cycling: Learning to estimate 3d human pose and shape by cycling between representations. In *AAAI*, pages 5561–5569, 2020. 8

[45] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018. 8

[46] Leonid Sigal, Alexandru Balan, and Michael J Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NeurIPS*, pages 1337–1344, 2008. 8

[47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, pages 1929–1958, 2014. 5

[48] Kristen Stubbs, Pamela J. Hinds, and David Wettergreen. Challenges to grounding in human-robot interaction. In *HRI*, pages 357–358, 2006. 1

[49] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, pages 5349–5358, 2019. 2, 5, 6, 8

[50] 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. 4

[51] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *CVPR*, pages 9661–9670, 2019. 8

[52] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *CVPR*, pages 195–204, 2019. 3, 8

[53] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NeurIPS*, pages 5242–5252, 2017. 8

[54] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*, 2017. 8

[55] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018. 8

[56] 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. 2, 5

[57] Zhe Wang, Daeyun Shin, and Charless C. Fowlkes. Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In *ECCV*, pages 523–540, 2020. 5

[58] Yuxin Wu and Kaiming He. Group Normalization. In *ECCV*, pages 3–19, 2018. 5

[59] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, pages 7760–7770, 2019. 1, 8

[60] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, pages 2148–2157, 2018. 8

[61] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3D human pose estimation. In *NeurIPS*, 2020. 5

[62] Zhenyu Zhang, Stephane Lathuiliere, Elisa Ricci, Nicu Sebe, Yan Yan, and Jian Yang. Online depth learning against forgetting in monocular videos. In *CVPR*, pages 4494–4503, 2020. 8