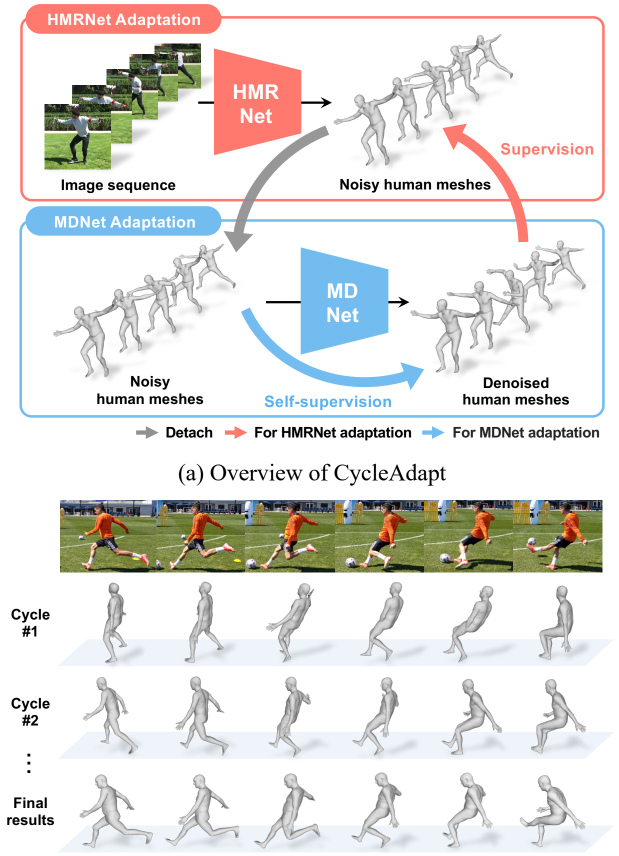
Cyclic Test-Time Adaptation on Monocular Video for 3D Human Mesh Reconstruction

Hyeongjin Nam[1,3] Daniel Sungho Jung[2,3] Yeonguk Oh[1,3] Kyoung Mu Lee[1,2,3]
[1]Dept. of ECE&ASRI, [2]IPAI, Seoul National University, Korea
[3]SNU-LG AI Research Center
{namhjsnu28,dqj5182,namepllet,kyoungmu}@snu.ac.kr

Abstract

Despite recent advances in 3D human mesh reconstruction, domain gap between training and test data is still a major challenge. Several prior works tackle the domain gap problem via test-time adaptation that fine-tunes a network relying on 2D evidence (e.g., 2D human keypoints) from test images. However, the high reliance on 2D evidence during adaptation causes two major issues. First, 2D evidence induces depth ambiguity, preventing the learning of accurate 3D human geometry. Second, 2D evidence is noisy or partially non-existent during test time, and such imperfect 2D evidence leads to erroneous adaptation. To overcome the above issues, we introduce CycleAdapt, which cyclically adapts two networks: a human mesh reconstruction network (HMRNet) and a human motion denoising network (MDNet), given a test video. In our framework, to alleviate high reliance on 2D evidence, we fully supervise HMRNet with generated 3D supervision targets by MDNet. Our cyclic adaptation scheme progressively elaborates the 3D supervision targets, which compensate for imperfect 2D evidence. As a result, our CycleAdapt achieves state-of-the-art performance compared to previous test-time adaptation methods. The codes are available in here.



(a) Overview of CycleAdapt

(b) Denoised results of MDNet as the cycle repeats

Figure 1. (a) We propose CycleAdapt that iteratively adapts the human mesh reconstruction network (HMRNet) and the human motion denoising network (MDNet) in a cyclic fashion. (b) As the cycle repeats, MDNet produces progressively accurate 3D human meshes as reliable 3D supervision targets for HMRNet, which in turn results in improved outputs of HMRNet.

1. Introduction

3D human mesh reconstruction (HMR) has gained popularity in many applications, such as AR/VR gaming, fitness tracking, and virtual try-on. Despite recent advances, one of the major bottlenecks is the prohibitive cost of collecting 3D training data on in-the-wild images, which are taken in our daily environments. Due to the challenge, most of HMR methods are commonly trained on Motion Capture (MoCap) [10, 29] datasets. While such datasets provide accurate 3D annotations obtained from sophisticated capturing devices, they contain limited human poses with less diverse image appearances compared to in-the-wild datasets. Accordingly, a domain gap arises in which performance in the test environment severely drops. In this work, we tackle the challenging domain gap problem via a test-time adaptation scheme that adapts a pre-trained HMR network to a given test in-the-wild video.

Most of the previous test-time adaptation methods [32, 7, 6, 40] fine-tune an HMR network via weak supervision with 2D evidence from test images, such as 2D human keypoints or silhouettes. They mainly rely on 2D reprojection loss that enforces the projection of reconstructed mesh to be close to the 2D evidence. However, the 2D reprojection loss causes two critical issues. First, the depth ambiguity of 2D evidence hinders learning accurate 3D geometry since innumerable points in 3D space correspond to the same 2D point of the 2D evidence. Second, 2D evidence for computing the 2D reprojection loss is often imperfect at test time, which results in erroneous adaptation. While several previous methods [7, 6] assume that ground-truths (GTs) of 2D evidence are available at test time, it is far from the practical scenario. During the test time, since we cannot acquire GT 2D evidence, the 2D evidence should be estimated from test images for the adaptation. Accordingly, the 2D evidence contains estimation error and is even partially non-existent, especially under human truncations and occlusions. Such imperfect 2D evidence leads to erroneous adaptation, making the HMR network to produce inadequate reconstructions, as shown in Figure 2.

To overcome the above limitations, we propose CycleAdapt, a novel test-time adaptation framework for 3D human mesh reconstruction. Our framework consists of two networks: a human mesh reconstruction network (HMRNet) and a human motion denoising network (MDNet), as shown in Figure 1(a). Given a test video, these two networks are adapted on the test video in two stages: 1) HMRNet adaptation stage and 2) MDNet adaptation stage. In the HMRNet adaptation stage, HMRNet is fully supervised with 3D supervision targets generated from the MDNet as well as the 2D evidence. Initially, HMRNet reconstructs a human mesh sequence from an image sequence of the test video. Then, the reconstructed human meshes are forwarded into MDNet, where the human meshes are refined via human motion denoising. The motion denoising effectively complements ambiguous parts (*e.g.*, occluded human part) that the HMRNet cannot infer from the image context. The refined meshes from MDNet act as 3D supervision targets during adaptation of HMRNet. Thus, the HMRNet is fully supervised with the generated 3D supervision targets, which alleviates the high reliance on 2D evidence in learning accurate 3D geometry of test images.

In the MDNet adaptation stage, MDNet is updated in a self-supervised manner with only noisy human meshes reconstructed from HMRNet. Adaptation for MDNet is crucial as the MDNet is pre-trained based on 3D labels of a MoCap dataset. Due to the restricted environment of the MoCap dataset, human motion distribution in the MoCap dataset is far from the distribution of test video, resulting in the degraded performance of MDNet. In this regard, we also perform adaptation for MDNet to improve the motion denoising performance in the test video. Since 3D human mesh GTs are unavailable during test time, we design the MDNet to be trainable in a self-supervised manner. In our design, random parts of noisy human meshes are masked, then the MDNet learns to reconstruct the masked parts of noisy human meshes. This self-supervised learning enhances denoising performance on the test video, despite only using noisy human meshes from HMRNet.

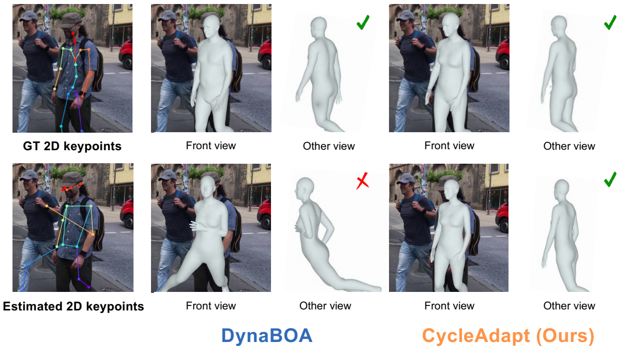


Figure 2. Given imperfect 2D evidence (keypoints) estimated from a test image, the previous test-time adaptation method [6] fails while our CycleAdapt produces accurate reconstruction results.

As shown in Figure 1 (a), the two adaptation stages iterate in a cyclic fashion. As the cycle repeats, the MDNet produces progressively reliable 3D supervision targets for HMRNet, as shown in Figure 1 (b). The progressively elaborated 3D supervision complements the imperfect 2D evidence of test images, preventing erroneous adaptation of HMRNet. As a result, our CycleAdapt produces far more accurate and natural human mesh reconstructions than previous methods, by resolving the major problems with the 2D evidence. We present an extensive evaluation of the proposed framework under various scenarios.

Our contributions can be summarized as follows.

- We present CycleAdapt, a novel test-time adaptation framework for 3D human mesh reconstruction to mitigate the domain gap between training and test data.
- We propose human motion denoising network, which generates 3D supervision targets to fully supervise the human mesh reconstruction network. Our cyclic adaptation strategy progressively elaborates the 3D supervision targets to prevent erroneous adaptation.
- We show that our CycleAdapt outperforms the previous state-of-the-art methods in various scenarios.

2. Related works

Domain adaptation for 3D human mesh reconstruction. Domain adaptation has recently emerged as a powerful strategy to alleviate the domain gap problem in 3D human mesh reconstruction. Joo *et al*. [13] proposed a method that fine-tunes a pre-trained network to the groundtruth 2D keypoints of target images. Mugaludi *et al*. [32] presented 2D silhouette-based supervision on adaptation for human mesh reconstruction network. Guan *et al*. [7] proposed BOA, an online adaptation framework with a bilevel optimization strategy to incorporate temporal consistency. Here, the training objective for the temporal consistency is computed based on the distance between predicted and target 2D joint coordinates. Guan *et al*. [6] further extended BOA into DynaBOA by introducing image retrieval and dynamic update strategy. Weng *et al*. [40] proposed to generate synthetic images and the corresponding human meshes, which are utilized in the adaptation.

The major difference of our CycleAdapt compared to prior works is that CycleAdapt generates 3D supervision targets corresponding to test images, to fully supervise the HMRNet during adaptation. BOA [7] and DynaBOA [6] construct 3D loss utilizing an external MoCap dataset [10] and apply the 3D loss for image samples from the MoCap dataset. Here, there is no 3D supervision for the test images during adaptation. Likewise, Weng *et al*. [40] also constructs 3D loss with their synthesized data, but only 2D reprojection loss is applied for the test images. On the other hand, CycleAdapt constructs 3D loss for test images, by using 3D supervision targets produced by MDNet. This 3D supervision is significantly helpful in learning accurate 3D geometry, where its effectiveness is provided in Section 5.2.

3D human mesh reconstruction. Most of the existing human mesh reconstruction methods [14, 34, 20, 19, 45, 30, 23, 21, 22, 4, 23] are based on parametric 3D human mesh model (*i.e.*, SMPL [25]), predicting parameters of the human mesh model. Kanazawa *et al*. [14] proposed an end-to-end trainable framework with adversarial loss to reconstruct plausible 3D human mesh. Pavlakos *et al*. [34] used 2D joint heatmaps and human silhouettes for accurate prediction of SMPL parameters. Kolotouros *et al*. [20] introduced a self-improving framework with an iterative fitting scheme. Kocabas *et al*. [19] proposed a part-guided attention mechanism for robustness on human occlusion. Zhang *et al*. [45] used mesh-aligned features to rectify SMPL parameter prediction. Moon *et al*. [30] utilized local and global image features for accurate human mesh reconstruction. Despite such advances in 3D human mesh reconstruction, the domain gap problem is still a major challenge, with a lack of studies on overcoming the discrepancy between training and test data.

Human motion denoising. Recent researches [27, 36, 46, 41, 43] have studied to leverage human motion prior to improve the reconstruction accuracy of 3D human meshes. Luo *et al*. [27] used a Variational Autoencoder (VAE) [17] to obtain coarse human motion for human motion estimation from a video. Rempe *et al*. [36] introduced test-time optimization for robust reconstruction from observation by leveraging a human motion generative model. Yuan *et al*. [41] proposed a method to infill missing human meshes from various occlusions. Zeng *et al*. [43] addressed varied estimation errors from a human mesh reconstruction network with an FCN-based denoising strategy. Zeng *et al*. [42] showed that reconstruction accuracy can be improved by completing removed human poses from 10% sampled video frames without any image context.

Different from all the above methods, we firstly address the test-time adaptation for human motion denoising. Existing motion denoising methods require GT human mesh sequences to learn the latent space of human motion generative model or supervise their predicted human motion. However, GT human mesh sequences are unavailable in the test-time adaptation scenario. Accordingly, we design the MDNet to be trainable without human mesh GTs, in a self-supervised manner. With self-supervised learning, MDNet is progressively adapted on the test domain in human motion, during the cyclic adaptation.

3. CycleAdapt

In the following sections, we first describe the overview of our cyclic adaptation framework, which consists of HMRNet and MDNet (Section 3.1). Then, we provide a detailed description for HMRNet adaptation and MDNet adaptation (Sections 3.2 and 3.3).

3.1. Cyclic adaptation

The main goal of CycleAdapt is fine-tuning two pretrained networks, HMRNet $\mathcal{M}_{\text{HMR}}$ and MDNet $\mathcal{M}_{\text{MD}}$, to enhance the reconstruction performance of HMRNet on a given test video $\mathbf{X}$. Algorithm 1 shows the overall adaptation procedure for HMRNet and MDNet. Each network outputs SMPL parameters $\{\theta, \beta\}$, then we can reconstruct 3D human mesh by forwarding the obtained parameters to the SMPL model [25]. The outputs of each network are temporally stored in a dictionary D for the effective adaptation, where D_i denotes intermediate outputs corresponding to ith frame of the test video. At the start of the algorithm, the dictionary D is initialized with dummy values, zero vectors. Next, HMRNet and MDNet are iteratively adapted with cycles $C = 12$.

A single cycle consists of two stages: 1) HMRNet adaptation stage and 2) MDNet adaptation stage. In the HMRNet adaptation stage, we sample ith image $\mathbf{x}_i$ from the test video and fetch ith SMPL parameters $\{\theta'_i, \beta'_i\}$ from the dictionary D. The HMRNet is updated by using fetched SMPL parameters as 3D supervision targets (Section 3.2). Then, we store the outputs $\{\hat{\theta}_i, \hat{\beta}_i\}$ of HMRNet in the dictionary D. In the MDNet adaptation stage, consecutive SMPL pose parameters $\{\hat{\theta}_j, \dots, \hat{\theta}_{j+T-1}\}$ are fetched from the dictio-

Algorithm 1 Pseudocode of Cyclic Adaptation

```
Input: Test frames $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$
Output: SMPL parameters $\{\hat{\theta}_i, \hat{\beta}_i\}_{i=1}^{N}$
Initialize dictionary $D$
for cycle $c = 1, \ldots, C$ do
    # HMRNet adaptation stage
    while sample $\mathbf{x}_i \sim \mathbf{X}$ do
        $\{\theta'_i, \beta'_i\} \leftarrow D_i$ # pseudo-GTs from previous cycle
        $\{\hat{\theta}_i, \hat{\beta}_i\} \leftarrow \mathcal{M}_{\text{HMR}}(\mathbf{x}_i)$
        Update $\mathcal{M}_{\text{HMR}}$ with $L_{\text{HMR}}$
        $D_i \leftarrow \{\hat{\theta}_i, \hat{\beta}_i\}$
    end while
    # MDNet adaptation stage
    while sample $\{\hat{\theta}_j, \ldots, \hat{\theta}_{j+T-1}\} \sim D$ do
        $\{\hat{\theta}'_j, \ldots, \hat{\theta}'_{j+T-1}\} \leftarrow \mathcal{M}_{\text{MD}}(\hat{\theta}_j, \ldots, \hat{\theta}_{j+T-1})$
        Update $\mathcal{M}_{\text{MD}}$ with $L_{\text{MD}}$
        $D_j, \ldots, D_{j+T-1} \leftarrow \{\hat{\theta}'_j, \ldots, \hat{\theta}'_{j+T-1}\}$
    end while
end for
```

nary D based on a randomly sampled frame index j, where $T = 49$ denotes the length of the sequence. The MDNet is updated based on a self-supervised learning scheme that only employs the fetched SMPL pose parameters (Section 3.3). Then, we store the outputs $\{\hat{\theta}'_j, \ldots, \hat{\theta}'_{j+T-1}\}$ of MDNet in the dictionary D, and the stored outputs are utilized for HMRNet adaptation stage in the next cycle. The detailed pipeline of a single cycle is illustrated in Figure 3. In the following sections, the frame index notations i and j will be omitted for simplicity.

3.2. HMRNet adaptation stage

The HMRNet $\mathcal{M}_{\text{HMR}}$ takes each single image $\mathbf{x} \in \mathbb{R}^{3\times224\times224}$ of a test video and predicts the pose parameters $\hat{\theta} \in \mathbb{R}^{144}$, shape parameters $\hat{\beta} \in \mathbb{R}^{10}$, and camera parameters $\hat{\mathbf{k}} \in \mathbb{R}^{3}$. By forwarding the predicted parameters $\{\hat{\theta}, \hat{\beta}\}$ to the SMPL model, the 3D human mesh coordinates $\hat{\mathbf{M}} \in \mathbb{R}^{6890\times3}$ are obtained. For HMRNet, we use ResNet-50 [9] as a backbone to extract an image feature from the input image after removing the fully-connected layer of the last part of the original ResNet. Then, we attach three fully-connected layers to regress SMPL parameters from the image feature, following Kanazawa *et al*. [14]. The HMRNet is pre-trained on a source dataset containing accurate 3D human labels, such as MoCap dataset [10] and synthetic dataset [38]. For the pre-training, we follow the conventional scheme of 3D human mesh reconstruction [20].

To adapt the HMRNet, we fetch the SMPL parameters $\{\theta', \beta'\}$, which are produced by MDNet in the previous cycle, from the dictionary D. We use the fetched SMPL parameters as 3D supervision targets to supervise predictions of HMRNet. Based on the 3D supervision targets, HMRNet

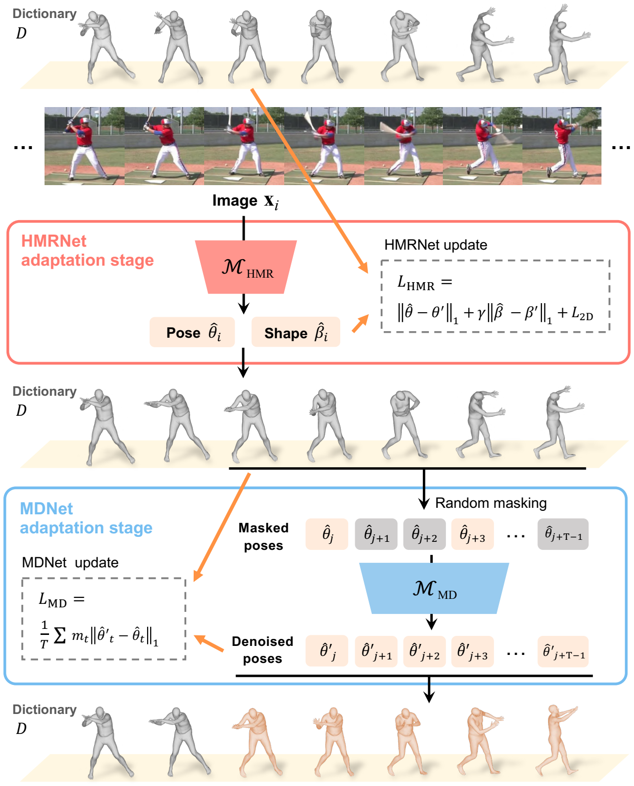

Dictionary
D
...
...
Image $\mathbf{x}_i$
HMRNet adaptation stage
$\mathcal{M}_{\text{HMR}}$
HMRNet update
$L_{\text{HMR}} =$
$\|\hat{\theta} - \theta'\|_1 + \gamma\|\hat{\beta} - \beta'\|_1 + L_{2\text{D}}$
Pose $\hat{\theta}_i$
Shape $\hat{\beta}_i$
Dictionary
D
Random masking
MDNet adaptation stage
Masked poses
$\hat{\theta}_j$ $\hat{\theta}_{j+1}$ $\hat{\theta}_{j+2}$ $\hat{\theta}_{j+3}$ ... $\hat{\theta}_{j+T-1}$
MDNet update
$L_{\text{MD}} =$
$\frac{1}{T}\sum m_t\|\hat{\theta}'_t - \hat{\theta}_t\|_1$
$\mathcal{M}_{\text{MD}}$
Denoised poses
θ'_j θ'_{j+1} θ'_{j+2} θ'_{j+3} ... $\hat{\theta}'_{j+T-1}$
Dictionary
D

Figure 3. The pipeline of a single cycle of CycleAdapt. In the HMRNet adaptation stage, HMRNet is adapted based on outputs of MDNet from the previous cycle. In the MDNet adaptation stage, MDNet is adapted in a self-supervised manner by only using outputs of HMRNet.

is adapted by minimizing the loss function L_{HMR} as follows:

$$L_{\text{HMR}} = L_{\text{SMPL}} + L_{2\text{D}}. \tag{1}$$

L_{SMPL} computes the L1 distance between predicted SMPL parameters and outputs of MDNet from the previous cycle as follows:

$$L_{\text{SMPL}} = \|\hat{\theta} - \theta'\|_1 + \gamma\|\hat{\beta} - \beta'\|_1, \tag{2}$$

where $\gamma = 0.001$. In the $c = 1$ cycle, L_{SMPL} is set to 0 since there are no stored outputs of MDNet in the dictionary. $L_{2\text{D}}$ is 2D reprojection loss that enforces the projection of reconstructed human mesh to be close to the 2D human keypoints, as follows:

$$L_{2\text{D}} = \|\mathbf{\Pi}_{\hat{\mathbf{k}}}(\mathcal{J}\hat{\mathbf{M}}) - \mathbf{J}^{2\text{D}}\|_1, \tag{3}$$

where $\mathbf{\Pi}(\cdot)$, $\mathcal{J}$, and $\mathbf{J}^{2\text{D}}$ denote a projection function, a joint regression matrix, and 2D keypoints predicted by an off-the-shelf 2D human pose estimator [2], respectively. The projection function $\mathbf{\Pi}(\cdot)$ performs weak-perspective projection based on the predicted camera parameters $\hat{\mathbf{k}}$.

3.3. MDNet adaptation stage

The MDNet $\mathcal{M}_{\text{MD}}$ takes a sequence of SMPL pose parameters $\{\hat{\theta}_0, \ldots, \hat{\theta}_{T-1}\}$ predicted from HMRNet and produces denoised pose parameters $\{\hat{\theta}'_0, \ldots, \hat{\theta}'_{T-1}\}$ toward natural human motion. We design MDNet by stacking multiple fully-connected layers with layer normalization. MDNet is pre-trained on a source dataset, a MoCap dataset [10], which contains 3D labels of human motions. For the pre-training, we first synthesize noise from GT human meshes from the MoCap dataset [10] and train the MDNet with pairs of noisy and GT human meshes. Further detail of the network architecture and the pre-training scheme is provided in the supplementary material.

When adapting MDNet, the main issue is that there is no GT 3D label corresponding to the noisy SMPL pose parameters at the test time. In this regard, motivated by Davlin *et al.* [5] and He *et al.* [8], we leverage a self-supervised learning strategy based on masking. Given a sequence of noisy SMPL pose parameters $\{\hat{\theta}_0, \ldots, \hat{\theta}_{T-1}\}$, we randomly mask half of the pose parameters $\lceil T/2 \rceil$ with zero vectors. Then, MDNet predicts the masked parts to make the entire pose sequence appear as a natural human motion. With only the noisy SMPL pose parameters, this strategy successfully learns human motion prior of the test video to improve the motion denoising performance. We describe its effectiveness in Section 5.2. The loss function for the MDNet adaptation is

$$L_{\text{MD}} = \frac{1}{T} \sum_{t=0}^{T-1} m_t \|\hat{\theta}'_t - \hat{\theta}_t\|_1, \tag{4}$$

where m_t denotes tth masking value that is set to one when the corresponding pose parameter is masked.

4. Implementation details

PyTorch [33] is used for implementation. The human body region is cropped using a GT bounding box for reconstructing 3D human mesh, following previous works [14, 20, 7]. When the bounding box is not available, an off-the-shelf human detector [35] is utilized for obtaining the bounding box. For all adaptation stages, weights of network are updated by Adam optimizer [16] with a mini-batch size of 32. An initial learning rate is set to 5×10^{-5} and reduced to 1×10^{-6} by a cosine annealing strategy [26]. A single NVIDIA GTX 2080 Ti GPU is used for all experiments.

5. Experiment

5.1. Datasets and evaluation metrics

Human3.6M. Human3.6M [10] is a large-scale MoCap dataset that is widely used in the 3D human mesh reconstruction community. Since this dataset is collected in a restricted environment with indoor setting, it lacks the diversity of human motions and image appearances. We use its training set as the source dataset, which is used for pre-training HMRNet and MDNet.

Evaluation networks	MPJPE	PA-MPJPE	MPVPE
HMRNet	98.7	59.8	112.3
MDNet before adaptation	114.2	62.6	134.4
MDNet after adaptation	**96.2**	**58.3**	**110.6**

Table 1. Effectiveness of MDNet adaptation on human motion denoising performance. During adaptation, we freeze the HMRNet and only train the MDNet.

SURREAL. SURREAL [38] is a synthetic dataset that contains diverse 3D human poses but contains artificial image appearances. We use its training set as the source dataset to pre-train HMRNet.

3DPW. 3DPW [39] is an in-the-wild dataset, mainly captured in outdoor environments, and it contains natural and diverse image appearances compared to MoCap and synthetic datasets. We use its test set as the target dataset for test-time adaptation.

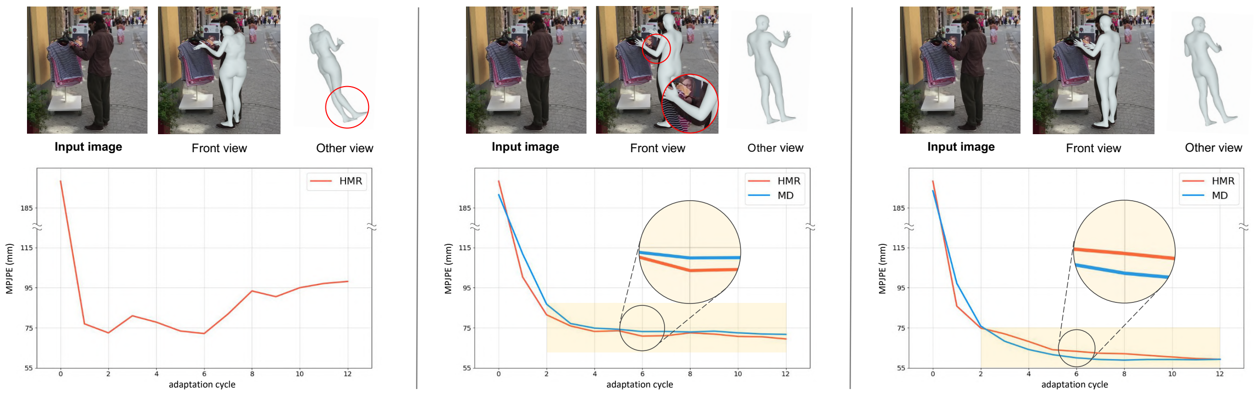
InstaVariety. InstaVariety [15] is an in-the-wild dataset, curated from Instagram videos. It contains numerous samples with dynamic human motions, such as basketball games and dancing. We use its test set as the target dataset for test-time adaptation. Since InstaVariety does not provide 3D GTs, we utilize it for qualitative comparisons only.

Evaluation metrics. For evaluation, we use the following metrics: (1) mean per joint position error (**MPJPE**), (2) Procrustes-aligned MPJPE (**PA-MPJPE**), (3) mean per vertex position error (**MPVPE**), and (4) acceleration error (**Accel**) that is used to measure temporal smoothness in video-based 3D human mesh reconstruction. All errors are measured in millimeters (mm) between the estimated and GT 3D coordinates after the root joint alignment.

5.2. Ablation study

We carry out ablation studies on test-time adaptation scenarios with Human3.6M [10] as source dataset and 3DPW [39] as target dataset. The 2D evidence (*i.e.*, 2D human keypoints) for adaptation is obtained via OpenPose [2].

Effect of MDNet adaptation on denoising performance. Table 1 shows that the MDNet adaptation improves motion denoising performance of MDNet, and the outputs of MDNet can act as reliable 3D supervision targets for HMRNet. In this ablation study, we only observe the effect on motion denoising performance while excluding the effect of HMRNet adaptation. To this end, we freeze HMRNet to provide fixed human mesh inputs for MDNet, with constant reconstruction accuracy (the first row). The MDNet before adaptation (the second row) shows inferior performance due to the domain gap caused by the difference in human motion distribution between the source dataset and the test video. On the other hand, MDNet after adaptation



(a) L_{2D} (b) $L_{\text{SMPL}} + L_{2D}$ (c) $L_{\text{SMPL}} + L_{2D}$ + Cyclic adapt. (**Ours**)

Figure 4. Comparison of qualitative results and MPJPE curves according to different adaptation strategies. We apply the adaptation on a 3DPW video sequence 'downtown_enterShop_00'.

Losses	Cyclic adapt.	MPJPE	PA-MPJPE	MPVPE
Base model (pre-trained on H36M)		230.3	123.4	253.4
* **Effectiveness of 3D supervision**				
L_{2D}	✗	125.5	74.4	154.0
$L^{\dagger}_{\text{SMPL}} + L_{2D}$	✗	115.2	68.5	142.0
$L_{\text{SMPL}} + L_{2D}$	✗	**96.9**	**60.7**	**114.5**
* **Effectiveness of cyclic adaptation**				
$L_{\text{SMPL}} + L_{2D}$	✗	96.9	60.7	114.5
$L_{\text{SMPL}} + L_{2D}$ **(Ours)**	✓	**87.7**	**53.9**	**105.7**

Table 2. Comparison of HMRNet's accuracy between different adaptation strategies. † denotes using Human3.6M [10] as external 3D dataset instead of using 3D supervision targets of MDNet.

(last row) achieves enhanced denoising performance by alleviating the domain gap.

Additionally, MDNet after adaptation also outperforms the HMRNet, which means the outputs of the MDNet can act as reliable 3D supervision targets for the HMRNet adaptation. While the HMRNet reconstructs 3D human meshes by focusing on the image context, the MDNet specializes in the temporal context of the human meshes for natural human motion. With the temporal context, the MDNet effectively complements ambiguous parts (*e.g.*, occluded human part) that the HMRNet cannot infer from the image context. Accordingly, the refined meshes provided by the MDNet act as beneficial 3D supervision targets during the adaptation of the HMRNet.

Effectiveness of 3D supervision by MDNet. The second block of Table 2 shows that adding 3D loss L_{SMPL} in the HMRNet adaptation stage (Section 3.2) significantly drops the errors compared to only using 2D reprojection loss L_{2D}. As shown in Figure 4, only using the 2D reprojection loss suffers from depth ambiguity, which results in improper reconstruction, especially in the depth direction.

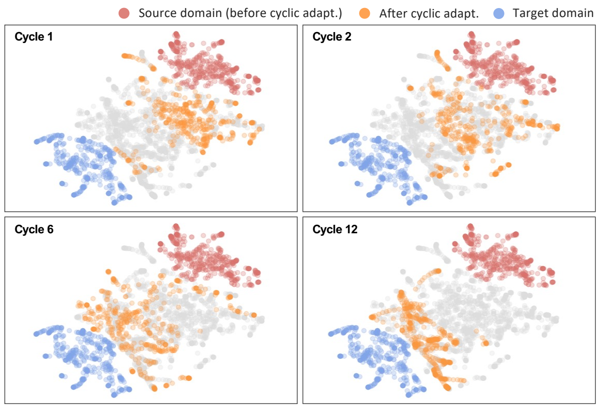


Figure 5. t-SNE visualization of image feature distribution during cyclic adaptation on a single test video. As the cycle progresses, the image feature distribution (orange) gets closer to the target domain distribution (blue).

On the one hand, we can enforce indirect 3D supervision as done by prior arts [7, 6, 40], training HMRNet with a mix-batch composed of test dataset and external 3D MoCap dataset [10]. In this strategy, the 3D loss $L^{\dagger}_{\text{SMPL}}$ is enforced only for samples from the external 3D dataset, without 3D supervision for test samples. Different from prior arts, we construct 3D loss L_{SMPL} for the test samples, by using the outputs of MDNet as 3D supervision targets. In our strategy, the HMRNet is fully supervised with the 3D loss L_{SMPL} for test samples. As shown in the second block of Table 2, our approach that enforces 3D supervision by MDNet significantly surpasses the prior strategies without using any external dataset for the test-time adaptation.

Effectiveness of cyclic adaptation. The last block of Table 2 shows that our cyclic adaptation strategy, which iteratively updates HMRNet and MDNet in a cyclic fashion, significantly boosts the performance of HMRNet. Here, the

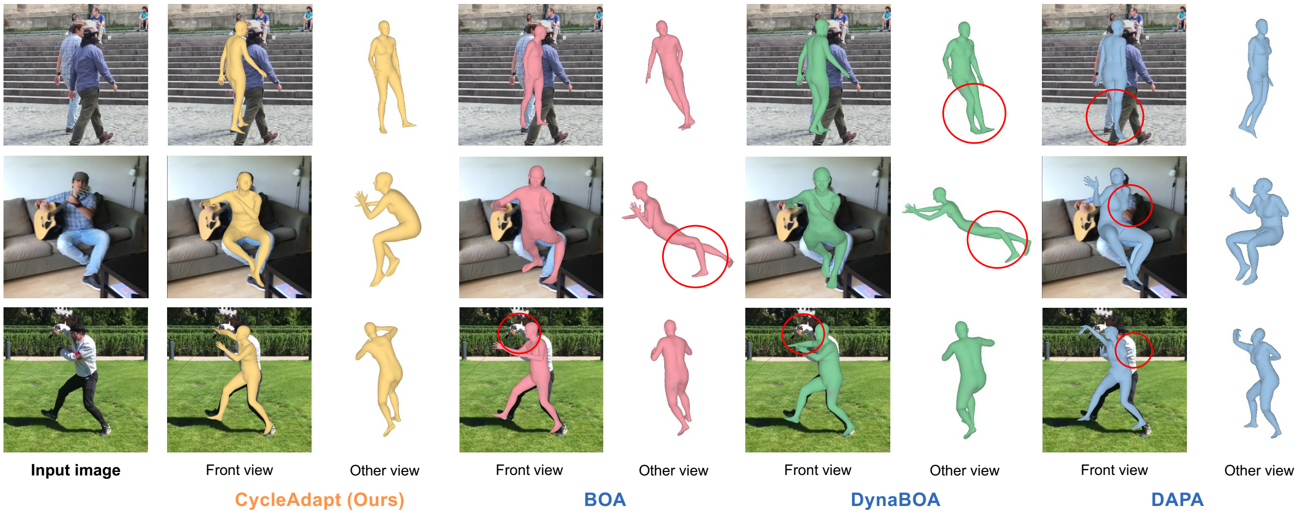


Figure 6. Qualitative comparisons with BOA [7], DynaBOA [6], and DAPA [40], when using Human3.6M [10] as source dataset and 3DPW [39] as target dataset. OpenPose [2] is used for all adaptations to obtain 2D human keypoints of test images. We highlighted their representative failure cases with red circles.

Motion denoising methods	MPJPE	PA-MPJPE	MPVPE
Gaussian 1D filter	92.0	57.5	108.1
Motion infiller [41]	92.4	55.5	109.1
SmoothNet [43]	92.5	54.8	112.1
MDNet (Ours)	**87.7**	**53.8**	**105.7**

Table 3. Comparison of HMRNet's accuracy according to different motion denoising methods used for the adaptation.

2D pose estimators	Methods	MPJPE	PA-MPJPE	MPVPE
Base model (pre-trained on H36M)		230.3	123.4	253.4
OpenPose [2]	BOA [7]	137.6	76.2	171.8
	DynaBOA [6]	135.1	73.0	168.2
	DAPA [40]	108.0	67.5	129.8
	CycleAdapt	**87.7**	**53.8**	**105.7**
HRNetw32 [37]	BOA [7]	139.5	79.9	172.1
	DynaBOA [6]	144.9	79.1	173.8
	DAPA [40]	104.2	66.9	128.0
	CycleAdapt	**86.9**	**53.2**	**102.6**
HRNetw32[37] + DarkPose [44]	BOA [7]	138.8	78.7	170.2
	DynaBOA [6]	142.0	77.3	170.0
	DAPA [40]	103.2	65.3	125.4
	CycleAdapt	**85.8**	**53.9**	**102.1**
GT	BOA [7]	73.2	46.2	91.4
	DynaBOA [6]	65.5	40.4	82.0
	DAPA [40]	75.0	46.5	92.4
	CycleAdapt	**64.7**	**39.9**	**76.7**

Table 4. Comparison of HMRNet's accuracy between different test-time adaptation methods, when using Human3.6M [10] as source dataset and 3DPW [39] as target dataset.

case of not performing cyclic adaptation indicates that only HMRNet is updated while MDNet is freezed during adaptation. When only adapting HMRNet, the error curve of MDNet is above that of HMRNet, as shown in Figure 4 (b). On the other hand, the MDNet with cyclic adaptation surpasses HMRNet after a few cycles, as shown in Figure 4 (c). Such MDNet consistently provides improved supervision targets for the next HMRNet adaptation stage. Then, the HMRNet after HMRNet adaptation stage produces more accurate human mesh reconstructions, which in turn, serves as better source of self-supervision in the next MDNet adaptation stage. As a consequence, this cyclic adaptation strategy progressively elaborates supervision targets for HMRNet, leading to the superior performance of HMRNet.

Figure 5 visualizes t-SNE, which shows that our cyclic adaptation effectively shifts the distribution of image features toward target domain. The image features are taken from the outputs of ResNet-50 [9] in the HMRNet. We performed t-SNE once with set of the image features from all cycles ($c = 1, 2, 6, 12$) and represented them with gray dots. The red and blue colors indicate the distribution when HMRNet is trained only on source dataset (*i.e.*, Human3.6M) and target dataset (*i.e.*, 3DPW), respectively. The orange color represents the distribution after a certain number of cycles. As shown in the change of orange dots, our cyclic adaptation framework effectively shifts the distribution of image features from the source domain (in red) toward the target domain (in blue), alleviating the domain gap.

Comparison with existing motion denoising methods. Table 3 shows the effectiveness of MDNet compared to

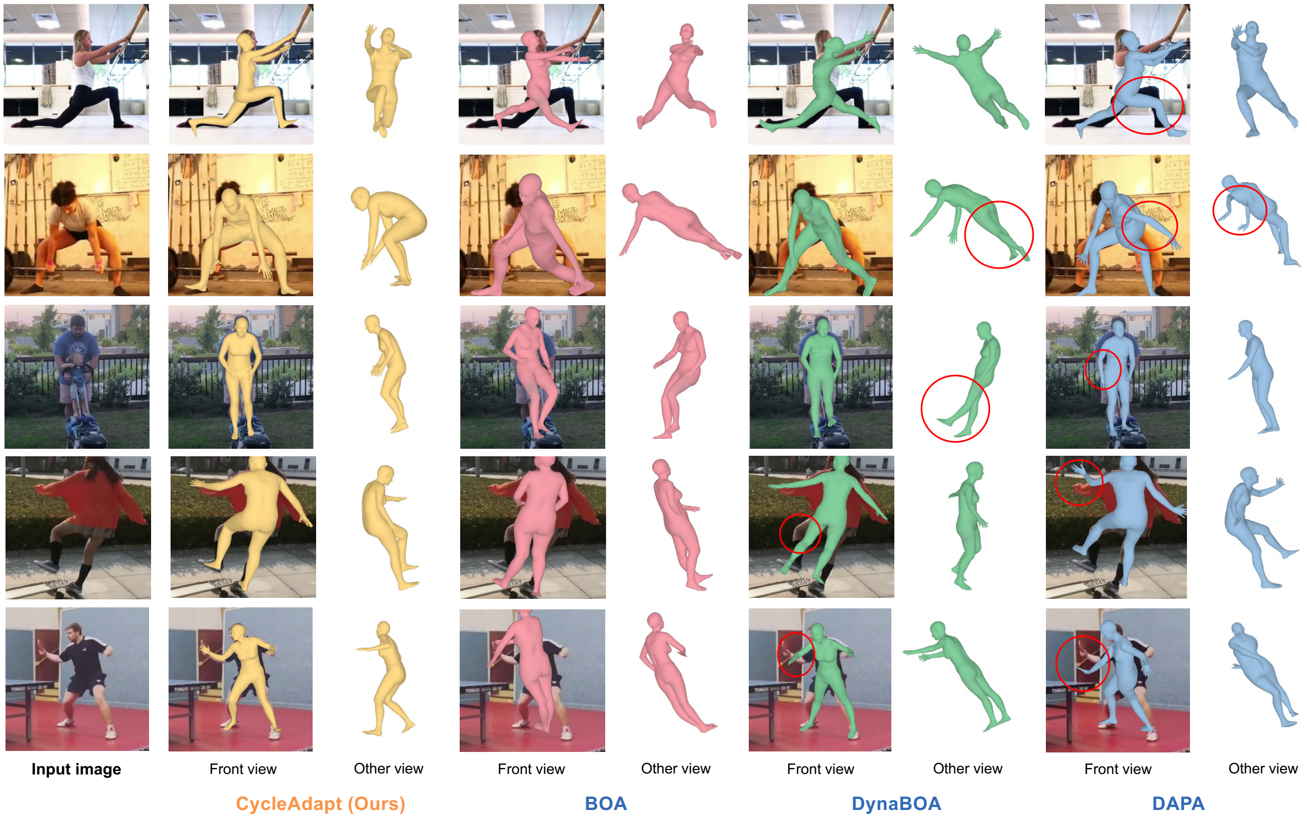


Figure 7. Qualitative comparisons with BOA [7], DynaBOA [6], and DAPA [40], when using Human3.6M [10] as the source dataset and InstaVariety [15] as the target dataset. OpenPose [2] is used for all adaptations to obtain 2D human keypoints of test images. We highlighted their representative failure cases with red circles.

existing human motion denoising methods in the test-time adaptation. Motion infiller [41] leverages a conditional variational autoencoder (CVAE) [17] trained on a large-scale MoCap dataset [28] with GT human mesh sequences. SmoothNet [43] is trained to minimize the distance between noisy and GT human mesh sequences. Different from the previous methods, MDNet is trainable without GT human meshes during test time. With the self-supervised learning scheme in Section 3.3, we can adapt MDNet to improve denoising performance on the test video. Therefore, our MDNet is more appropriate for providing elaborated supervision targets for HMRNet adaptation.

5.3. Comparison with state-of-the-art methods

We compare our CycleAdapt with recent test-time adaptation methods [7, 6, 40] for 3D human mesh reconstruction: BOA [7], DynaBOA [6], and DAPA [40]. Since all methods require 2D human keypoints of test images for adaptation, we obtain the 2D keypoints by using off-the-shelf 2D pose estimators [2, 37, 44]. All of their results are obtained with their officially released codes, and pre-trained HMRNet weights are equally set for a fair comparison.

Qualitative results. Figures 6 and 7 show that our CycleAdapt produces much better reconstruction results than the state-of-the-art test-time adaptation methods. In this comparison, we use Human3.6M [10] as source dataset to pre-train the HMRNet. Previous methods highly rely on 2D evidence from test images, which results in undesirable reconstruction results, especially in the depth direction. Furthermore, the projection alignment is often incorrect, caused by imperfect 2D evidence. Our CycleAdapt effectively resolves the high reliance problem on 2D evidence, which significantly benefits HMRNet to adapt on

Methods	MPJPE	PA-MPJPE	MPVPE
Base model (pre-trained on SURR)	193.2	92.0	216.5
BOA [7]	102.5	61.7	124.7
DynaBOA [6]	109.8	62.4	139.9
DAPA [40]	96.6	61.7	122.8
CycleAdapt (Ours)	**84.4**	**51.1**	**99.9**

Table 5. Comparison between different test-time adaptation methods, when using SURREAL [38] as the source dataset and 3DPW [39] as the target dataset. OpenPose [2] is used to obtain 2D human keypoints from test images for the adaptation.

	Methods	MPJPE	PA-MPJPE	MPVPE	Accel
image-based	HMR [14]	130.0	76.7	-	37.4
	SPIN [20]	96.9	59.2	116.4	29.8
	I2L-MeshNet [31]	93.2	57.7	110.1	30.9
	PyMAF [45]	92.8	58.9	110.1	-
	Pose2Pose [30]	**86.6**	54.4	**103.8**	16.2
	CycleAdapt (HMRNet)	87.7	**53.8**	105.7	**12.0**
video-based	HMMR [15]	116.5	72.6	139.3	15.2
	VIBE [18]	93.5	56.5	113.4	27.1
	TCMR [3]	95.0	55.8	111.3	6.7
	SmoothNet [43]	97.8	61.2	111.5	7.4
	CycleAdapt (MDNet)	**87.7**	**53.7**	**105.9**	**5.9**

Table 6. Comparison with existing 3D human mesh reconstruction methods. Our CycleAdapt achieves state-of-the-art performance by adapting networks pre-trained on Human3.6M [10], whereas other methods employ numerous datasets for the training.

test data. These qualitative results are consistent with the ablation study.

Quantitative results. Table 4 shows that our CycleAdapt achieves the best accuracy compared to the previous methods with various 2D pose estimators [2, 37, 44]. In this comparison, we use MoCap dataset (*i.e.*, Human3.6M [10]) as source dataset and 3DPW [39] as target dataset for test-time adaptation. The last block of Table 4 shows a scenario of using GT 2D human keypoints from test images, as done in BOA [7] and DynaBOA [40]. However, in practice, the GT 2D human keypoints are unavailable during test time. Accordingly, we cover a more practical scenario, using 2D pose estimators to obtain 2D human keypoints from test images. In the practical scenario, our CycleAdapt significantly outperforms previous methods with the same tendency in diverse 2D pose estimators, as shown in Table 4. Additionally, Table 5 shows the superior performance of CycleAdapt when using a synthetic dataset (*i.e.*, SURREAL [38]) as source dataset and 3DPW [39] as target dataset.

Table 6 shows that our CycleAdapt achieves state-of-the-art performance in 3D human mesh reconstruction, compared to both image- and video-based approaches. We compare the HMRNet with image-based networks and the MDNet with video-based networks, considering the type of network input. The compared 3D human mesh reconstruction methods exploit numerous training datasets [10, 29, 24, 1, 11, 12], to train their HMR networks. Despite using much less training data in pre-training, our CycleAdapt can achieve state-of-the-art performance by adaptation on the test dataset.

Running time. Table 7 shows that our CycleAdapt takes the shortest computational time during adaptation, compared to previous test-time adaptation methods. The running time is measured in the same environment with Intel Xeon Gold 6248R CPU and NVIDIA GTX 2080 Ti GPU, excluding pre-processing stages, such as pre-training and 2D pose estimation. For the measurement on the previous methods, we followed the same experimental setting from each method. BOA [7] and DynaBOA [6] demand a much longer time because there are two network update steps in their bilevel optimization algorithm for every single image. DAPA [40] also suffers from substantial adaptation time as it contains a rendering pipeline that generates a synthetic image for each test image, during adaptation. In contrast, our CycleAdapt takes much less time, although our framework additionally adapts MDNet along with HMRNet. As shown in Table 8, the MDNet adaptation stage requires minimal computational overhead and does not significantly affect the overall running time. Thus, our proposed framework has a significant advantage in running time.

BOA [7]	DynaBOA [6]	DAPA [40]	**CycleAdapt (Ours)**
840.3	1162.8	431.0	**74.1**

Table 7. Running time comparisons between different test-time adaptation methods, where the unit of time is millisecond (ms).

HMRNet adaptation stage	MDNet adaptation stage	Total
66.4	7.7	74.1

Table 8. Running time of each adaptation stage of our CycleAdapt, where the unit of time is millisecond (ms).

6. Conclusion

We propose CycleAdapt, a novel and powerful test-time adaptation framework for 3D human mesh reconstruction. Our framework addresses high reliance on 2D evidence of test images during adaptation, with the cyclic adaptation scheme that iteratively adapts a human mesh reconstruction network (HMRNet) and a human motion denoising network (MDNet) in a cyclic fashion. In our framework, the HMRNet is fully supervised with 3D supervision targets, which are outputs of the MDNet, as well as 2D evidence of test images. The 3D supervision targets are progressively elaborated by our cyclic adaptation strategy, which compensates for the imperfect 2D evidence, to prevent erroneous adaptation. We show that CycleAdapt significantly outperforms previous methods in various scenarios, both qualitatively and quantitatively.

Acknowledgements. This work was supported in part by the IITP grants [No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University), No. 2021-0-02068, and No.2023-0-00156], the NRF grant [No. 2021M3A9E4080782] funded by the Korea government (MSIT), and the SNU-LG AI Research Center.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017.

[3] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *CVPR*, 2021.

[4] Hongsuk Choi, Hyeongjin Nam, Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. Rethinking self-supervised visual representation learning in pre-training for 3D human pose and shape estimation. In *ICLR*, 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[6] Shanyan Guan, Jingwei Xu, Michelle Z He, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *TPAMI*, 2022.

[7] Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *CVPR*, 2021.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 2014.

[11] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[12] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[13] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation. In *3DV*, 2021.

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[15] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[19] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021.

[20] 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*, 2019.

[21] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

[22] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *CVPR*, 2021.

[23] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015.

[26] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *ICLR*, 2017.

[27] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3D human motion estimation via motion compression and refinement. In *ACCV*, 2020.

[28] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019.

[29] 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*, 2017.

[30] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPRW*, 2022.

[31] 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.

[32] Ramesha Rakesh Mugaludi, Jogendra Nath Kundu, Varun Jampani, et al. Aligning silhouette topology for self-adaptive 3D human pose recovery. In *NeurIPS*, 2021.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.

[34] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018.

[35] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[36] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. HuMoR: 3D human motion model for robust pose estimation. In *ICCV*, 2021.

[37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[38] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[39] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018.

[40] Zhenzhen Weng, Kuan-Chieh Wang, Angjoo Kanazawa, and Serena Yeung. Domain adaptive 3D pose augmentation for in-the-wild human mesh recovery. In *3DV*, 2022.

[41] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022.

[42] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. DeciWatch: A simple baseline for 10× efficient 2D and 3D pose estimation. In *ECCV*, 2022.

[43] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. SmoothNet: A plug-and-play network for refining human poses in videos. In *ECCV*, 2022.

[44] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020.

[45] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021.

[46] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4D human body capture in 3D scenes. In *ICCV*, 2021.