Table R1

Method	PCK	AUC	MPJPE	Time[s]
Vanilla-S	82.5	47.6	94.1	0.244
Vanilla-B	83.7	48.4	91.3	0.083
Online-S	83.6	48.2	92.2	0.027
Online-B	84.1	48.7	90.5	0.007

Table R2

Method	MPJPE
Baseline	45.0
Joint	42.6
ISO	41.8

Table R3

Method	PCK	AUC	MPJPE
[7]	65.3	33.4	135.7
[9]	75.9	36.3	-
[35]	62.4	30.9	147.4
Ours	83.6	48.2	92.2

Table R4

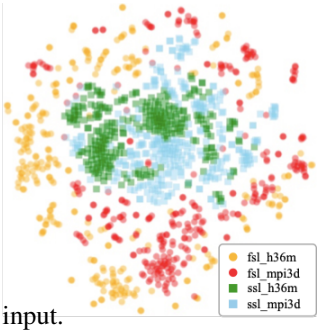
Method	PCK	AUC	MPJPE
[14]	60.9	30.2	150.1
[7]	65.3	33.4	135.7
Ours	83.6	48.2	92.2

To R1 **Q1. Batch adaptation.** ISO can perform model adaptation with a batch of instances, if available. See in Table R1, where **B** and **S** denote batch and single-instance ISO. Batch ISO outperforms single-instance ISO in both accuracy (from 94.1 to 91.3 in MPJPE for *Vanilla*) and efficiency ($\sim$3$\times$ and 4$\times$ faster for *Vanilla* and *Online*).
Q2. Comparison based on faster version of ISO will be added in the revision.
Q3. The pose discriminator is pretrained on the source data. During inference, it is updated using Eqn. (3) for each new target instance. To alleviate potential training bias from a single instance, we applied horizontal flip augmentation and the augmentation strategy (see Line 159-161) to update the discriminator. We will add more details in the revision.
Q4. Ablation studies on w/ and w/o ISO over H36M are given in Table R2. We observe *Joint* improves *Baseline* by a large margin. ISO makes marginal improvement over *Joint* (from 42.6 to 41.8 in MPJPE) since training and testing distribution are similar. However, even though there is no significant distribution shift, ISO still makes positive effect.

To R2 **Q1. Differences between ISO and model-fititng (MF) methods**: 1) The motivation and methodology are different. The MF methods (e.g., Kolotouros et al., CVPR 2019) aim to improve the model training by iterating regression (using a parametric human body model) and optimization. However, they do not consider how to generalize the model to new testing distributions that are different from the training ones. Our ISO focuses on mining distributional knowledge about the testing distribution from unlabeled instances via SSL and adapt the models accordingly to gain better generalizability. 2) The MF methods usually require a parametric model (e.g., SMPL) and a 3D body mesh initialization for model fitting; whereas ISO does not require these and thus is more general.
Q2. "Main improvements come from *Joint*" is only true for *Vanilla* as *Vanilla* is always re-initialized using *Joint* for each new instance. Thus, *Vanilla* cannot benefit the model too much as distributional information mined from a single instance is limited. When more distributional information is mined, ISO can make significant improvement. This can be observed from: 1) *Online* improves *Joint* by -4 in MPJPE (see Table S1). 2) Batch version of *Vanilla* improves the performance upon single-instance version of *Vanilla* by a large margin (see **Q1** in responses to **R1** and Table R1).
Q3.Differences from un-/weakly-supervised methods: un-/weakly-supervised methods are usually used for better training the model with more unlabeled *training data* (see Line 72-78) and rarely used in a transductive manner for testing; while ISO aims to improve the model's generalizability during inference via SSL. Among the suggested methods [7, 9, 35], [7] uses cycle-supervision and can be used in a transductive manner; while [9,35] requires multi-view data which are not available on 3DHP and 3DPW during testing. We compare ISO with them on 3DHP in Table R3. ISO significantly outperforms them. Comparison with them on 3DPW will be added in revision.
Q4. Qualitative results. We will visualize GT in figure 4 and show some results with noisy inputs in the revision.

To R3 **Q1. 2D pose input.** For fair comparison with [11, 46, 50], in Table 1, we use GT poses as inputs. However, ISO is robust to 2D pose noise as shown in Table 4 and thus can perform similarly well when using detected poses.
Q2. SSL methods under transductive learning (TL) setup. [A] cannot be evaluated under TL setup on 3DHP since it requires multi-view data while the test set only has single-view data. We re-implement other SSL methods [7, 14] and evaluate them on 3DHP under TL setup (see in Table R4). We can observe ISO outperforms them by a large margin.
Q3. Relation between ISO and TL methods will be added in the revision.

To R4 **Q1. Questions about model components.** To study these, we visualize the hidden features of different samples (random samples from H36M test set and samples from 3DHP) extracted by both heads using t-SNE (see right figure). The samples from 3DHP are the ones showing improvement after ISO, which present novel viewpoints and body sizes. We observe features from SSL head of both datasets (green & blue squares) are closer than the ones from FSL head (orange & red circles). Thus, SSL head learns features more robust to distribution change whereas FSL head learns more discriminative features from source data. The features from these two heads are thus disentangled. Shared feature extractor keeps these two kinds of features. FSL head would fail when facing unusual poses and viewpoints (e.g., top views) which are very ambiguous with only monocular 2D pose input.



Q2. Will online ISO overfit to the samples come at first? To answer this, we randomly shuffle the test set of 3DHP before performing online ISO. We conduct experiments for 8 times and obtain the statistics: PCK: 83.2 ± 0.43, AUC: 48.0 ± 0.30 and MPJPE: 92.9 ± 1.3. The small variance implies online ISO does not overfit to the sequential data. As discussed in **Q1**, the model overfitting to the SSL task would extract less discriminative features and thus hamper the performance. We will clarify this in the revision.