Uncertainty-Aware Adaptation for Self-Supervised 3D Human Pose Estimation

Jogendra Nath Kundu[1] Siddharth Seth[1*] Pradyumna YM[1*] Varun Jampani[2]
Anirban Chakraborty[1] R. Venkatesh Babu[1]
[1]Indian Institute of Science, Bangalore [2]Google Research

Abstract

The advances in monocular 3D human pose estimation are dominated by supervised techniques that require large-scale 2D/3D pose annotations. Such methods often behave erratically in the absence of any provision to discard unfamiliar out-of-distribution data. To this end, we cast the 3D human pose learning as an unsupervised domain adaptation problem. We introduce MRP-Net[1] that constitutes a common deep network backbone with two output heads subscribing to two diverse configurations; a) model-free joint localization and b) model-based parametric regression. Such a design allows us to derive suitable measures to quantify prediction uncertainty at both pose and joint level granularity. While supervising only on labeled synthetic samples, the adaptation process aims to minimize the uncertainty for the unlabeled target images while maximizing the same for an extreme out-of-distribution dataset (backgrounds). Alongside synthetic-to-real 3D pose adaptation, the joint-uncertainties allow expanding the adaptation to work on in-the-wild images even in the presence of occlusion and truncation scenarios. We present a comprehensive evaluation of the proposed approach and demonstrate state-of-the-art performance on benchmark datasets.

1. Introduction

3D human pose estimation forms a core component of several human-centric technologies such as augmented reality [24], gesture recognition [6], etc. Most of the 3D human pose estimation approaches heavily rely on fully supervised training objectives [15, 59, 80], demanding access to large-scale datasets with paired 3D pose annotation. However, the inconvenience of 3D pose acquisition stands as a significant bottleneck. Unlike a 2D pose, it is difficult to manually annotate an anthropomorphically constrained 3D pose for an in-the-wild RGB image. Thus, most of the paired 3D pose datasets are collected in lab environments via body-worn sensors or multi-camera studio setups [28, 78] that are difficult to install outdoors. This often limits the dataset diversity in terms of the variety in poses, appearances (background and lighting conditions), and outfits.

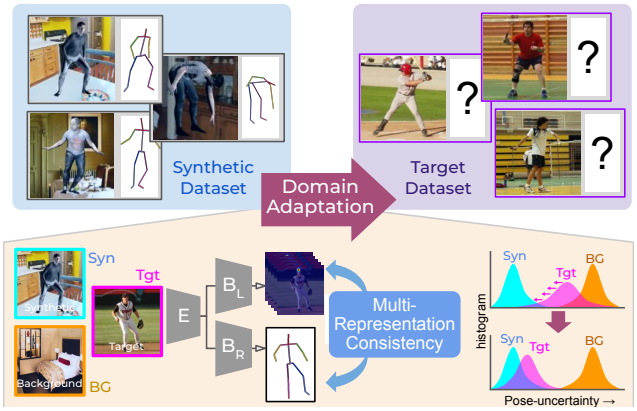

Figure 1. The proposed unsupervised adaptation framework utilizes a multi-representation consistency based uncertainty estimation for simultaneous OOD detection and adaptation.

Some works [69, 71] propose weakly supervised techniques to bypass the requirement of 3D pose annotations. Several of these works leverage available paired 2D pose datasets or off-the-shelf image-to-2D pose estimation networks [34, 63, 83]. To address the inherent 2D-to-3D ambiguity, some works either rely on multi-view image pairs [11, 32, 38] or utilize unpaired 3D pose samples [10, 88]. Though such methods perform well when evaluated on the same dataset, they lack cross-dataset generalization.

Consider a scenario where we want to deploy a 3D human pose estimation system in a new application environment (*i.e.* target domain). From the vendor's perspective, a general approach would be to improve the system's generalization via supervised training on a wide variety of labeled source domains [48]. However, target-specific training usually achieves the best performance beyond the generic system. Though it is not convenient to collect annotations for every novel deployment scenario, an effective unsupervised adaptation framework stands as the most practical way forward. Unsupervised adaptation [19, 37, 84] seeks a learning technique that can minimize the domain discrepancy be-

*equal contribution.
[1]Project page: https://sites.google.com/view/mrp-net

Table 1. Comparison of positive (in green) and negative (in red) attributes of ours against prior 3D human pose estimation methods.

Methods	Real Sup.			Synthetic 3D pose Sup.	Generalization capability	
	Multi view	2D pose	3D pose		Occlusion	Uncertainty
Zhou *et al*. [102]	✗	✓	✗	✗	✗	✓
Rhodin *et al*. [70]	✓	✗	✗	✗	✗	✗
Iqbal *et al*. [29]	✓	✓	✗	✗	✗	✗
Doersch *et al*. [17]	✗	✗	✗	✓	✗	✗
LCR-Net++ [74]	✗	✓	✓	✗	✓	✗
PoseNet3D [81]	✗	✓	✗	✗	✗	✗
Ours	✗	✗	✗	✓	✓	✓

tween a labeled source and an unlabeled target. Thus, the vendor has to collect unlabeled RGB inputs from the new environment to enable the adaptation process. Let us consider a different scenario where the target environment is identical to one of the source domains implying no domain-shift. Here, the vendor can choose to directly deploy the generic system without adaptation training. However, the system must have a provision to detect whether it is required to run the adaptation process. In other words, it should have the ability to discern out-of-distribution (OOD) scenarios [26,47,51]. Such an ability is more crucial while deploying in a continually changing environment [89], *e.g*. a model adapted for sunny weather conditions would fail while encountering rainy weather, thus requiring re-adaptation.

We propose a novel domain adaptation (DA) framework, *MRP-Net* (Fig. 1), equipped with uncertainty estimation [35] for the monocular 3D human pose estimation task. To this end, we use a multi-representation pose network with a common backbone followed by two pose estimation heads subscribing to two diverse output configurations; a) Heat-map based joint localization and b) Model-based parametric regression. This not only encourages ensemble-diversity required for uncertainty estimation [18] but also allows us to encompass the merits of both schools of thought [58, 80]. The former configuration advocates maintaining the spatial structure via a fully-convolutional design [56, 61, 79] while lacking provisions to inculcate structural articulation and bone-length priors. The latter advocates regressing a parametric form of the pose as a whole (via fully-connected layers) [30, 33, 49, 82] while allowing model-based structural prior infusion [39, 67]. We use the 3D graphics-based synthetic SURREAL dataset [86] as the labeled source domain to supervise our backbone network.

In addition, we derive useful measures to quantify the prediction uncertainty at two granularity levels; viz a) *pose-uncertainty*, b) *joint-uncertainty*. During training, we utilize both a labeled source and a dataset of backgrounds (BG) to elicit the desired behavior of the uncertainties. Here, the backgrounds approximate an extreme out-of-distribution scenario. Upon encountering the unlabeled target, the adaptation process seeks to reduce the target uncertainties alongside a progressive self-training on a set of reliable pseudo-labels. Alongside the adaptation for datasets with full-body visibility, the *joint-uncertainty* lays a suitable ground to expand our adaptation to work on in-the-wild target domain (unlabeled) with partial body visibility (*i.e*. under external occlusion or truncated frame scenarios). We present an extensive evaluation of the proposed framework under a variety of source-to-target settings. In summary:

- We propose a novel domain adaptation framework, *MRP-Net*, that uses a multi-representation pose network. Here, *pose-uncertainty* is quantified as the disagreement between pose predictions through the two output heads subscribing towards two diverse design configurations (model-free versus model-based).
- We propose to utilize negative samples (backgrounds and simulated synthetic joint-level occlusions) to improve the effectiveness of the proposed pose and joint uncertainties. The presence of negatives also helps to retain the uncertainty estimation ability even while adapting to a novel target scenario.
- Our synthetic (SURREAL) to in-studio adaptation outperforms the comparable prior-arts on Human3.6M [28]. Our in-studio (Human3.6M) to in-the-wild adaptation achieves state-of-the-art performance across four datasets. We show uncertainty-aware 3D pose estimation results for unsupervised adaptation to in-the-wild samples with partial body visibility.

2. Related Works

Table 1 shows a comparison of our approach against related prior approaches. Here, Sup. stands for supervision.

Domain Adaptation. Cao *et al*. [9] propose to apply discriminator based discrepancy minimization technique for the animal pose estimation task. To address the synthetic-to-real domain gap for 3D human pose estimation, Doersch *et al*. [17] propose to use optical-flow and 2D keypoints as the input as these representations are least affected by domain shift unlike RGB images (texture and lighting variations). Similarly, Zhang *et al*. [97] propose to leverage multi-modal input, such as depth and body segmentation masks. Mu *et al*. [60] leverage several consistency losses to effectively adapt from source to target. Our proposed framework does not access any such auxiliary input modality. Recently, some works [77, 96] propose online test-time adaptation of 3D human pose estimation from in-studio source to in-the-wild target.

Pose estimation in presence of occlusion. In literature, we find some methods that address human pose estimation in presence of partial occlusion. Several works design techniques to estimate location of the occluded keypoint conditioned on the unoccluded ones while accessing additional spatio-temporal [13, 14, 16, 68, 73] or scene related context [41,94,95]. Mehta *et al*. [55] propose to use occlusion-robust pose-maps to address partial occlusion scenarios.

Monocular 3D human pose estimation. In literature, we find two broad categories viz. a) methods that directly infer the 3D pose representation [1, 4, 75] and b) methods using model-based parametric representation [3, 5, 7, 40, 64]. The former directly maps the input image to the 3D pose while the latter maps images to latent parameters of a predefined parametric human model. The latter setup provides a suitable ground to impose the kinematic pose priors via adversarial training [36, 42]. The former setup is further categorized into one-stage [62, 65, 66, 80, 91, 100] and two-stage methods [27, 53, 59, 99]. One-stages approaches directly map images to the 3D poses. Whereas, two-stage methods first map images to an 2D pose representation followed by another mapping to perform the 2D-to-3D lifting.

Pose estimation via multi-head architecture. PoseNet3D [81] employs a student-teacher multi-head framework. However, the primary task is 2D-to-3D lifting where they rely on 2D pose predictions obtained from fully supervised image-to-2D pose model [61]. Unlike PoseNet3D, we do not leverage in-the-wild 2D pose annotations or temporal consistency. Further, prior arts [23, 72] also employ similar multi-head architecture to leverage auxiliary supervision or to improve predictions through consistency losses. To the best of our knowledge, none of the prior-arts utilize such architecture for OOD or self-adaptation to unlabeled target.

3. Approach

We aim to prepare a pose estimation network that can discern OOD samples by delivering a high prediction uncertainty for such inputs. Simultaneously, the network should not compromise on pose estimation performance for in-domain inputs. Sec. 3.1 first discusses the pros and cons of the two widely used design configurations specific to output representation of human pose estimation networks. We describe the key design components of the proposed *MRP-Net* architecture, following which we propose intuitive ways to quantify the pose and joint uncertainties. Sec. 3.2 illustrates the training procedure to progressively strengthen and leverage the *pose-uncertainty* for the unsupervised DA setting. In Sec 3.3, we leverage the *joint-uncertainties* as a means to expand the adaptation to a broader scope, *i.e.* to in-the-wild targets in the presence of occlusion and truncations.

3.1. Pose estimation architecture

In literature, pose estimation architectures employ one of the following two design configurations.

a) Localization-based representation. Most of the popular 2D pose estimation approaches employ fully convolutional architectures (such as hourglass networks [61]), where the final pose is realized via J heatmaps, one for each joint [56]. Here, the heatmaps are treated as spatial probability distributions (PDFs) with a probability peak near the spatial joint location. This can also be viewed as a localization based model-free design as it refrains from utilizing the joint-connectivity and the bone-length knowledge.

b) Regression-based representation. Here, networks aim to directly regress joint coordinates or some rich parametric representations (latent) of the final pose [30, 82]. Networks employ fully-connected layers after the back-bone CNN, thereby breaking away from the spatial structure to learn a highly non-linear mapping, unlike the localization-based design. One can easily inculcate joint-connectivity or bone-length priors via model-based design with integrated forward kinematics [44, 67, 101]. However, such model-based representation does not allow a provision to extract joint-level uncertainty as it sees the pose as a whole.

Normally trained systems often behave erratically in the absence of any provision to discard out-of-distribution inputs. In literature, ensemble-based systems [45] have been used to derive useful uncertainty measures. Several approaches resort to random initialization or dataset bootstrapping to induce *ensemble-diversity* which is crucial to realize a robust uncertainty quantification metric.

3.1.1 MRP-Net architecture

Keeping in mind the computational overhead of full network ensembles, we decide to develop multi-head ensembles with a common CNN backbone. As shown in Fig. 2, the multi-head ensemble consists of a common encoder backbone E which is followed by two ensemble heads that are denoted as B_L and B_R. Unlike the random initialization strategy, we propose to maintain ensemble diversity by following the above discussed pose modeling configurations.

a) Joint-localization at B_L output. The localization branch B_L is a convolutional decoder which outputs heatmap PDFs, $\tilde{h} : \{\tilde{h}^{(j)}\}_{j=1}^{J}$ (via spatial-softmax). These heatmaps are converted to 2D joint coordinates via a soft-argmax operation, $\tilde{q}^{(j)} = \sum_v (v) h^{(j)}(v)$. Here, $v : [v_x, v_y]$ denotes the spatial grid index. We also extract joint-confidence, $\tilde{w}$ as $\tilde{w}^{(j)} = \max_v h^{(j)}(v)$.

b) Kinematic-parameterization at B_R output. On the other hand, the regression branch B_R consists of several fully-connected layers to regress a 3D pose parameterization, $\hat{p}^l$ and camera-parameters, $\hat{c}$. We design a simple kinematic model based on the knowledge of hierarchical limb connectivity and relative bone-length ratios. We aim to disentangle the rigid camera variations (in camera space) from the non-rigid limb articulations (in canonical space). Here, the non-rigid articulations are modeled at the view-independent canonical space. Note that, in canonical space, the pelvis joint exactly aligns with the origin while the skeleton faces towards the positive X-axis, thus making it a view-independent pose representation, $\hat{p}^c \in \mathbb{R}^{J\times 3}$. In our convention, the skeleton-face is obtained as the cross-product direction of two vectors; i.e., left-hip to neck and left-hip to right-hip. However, directly regressing the

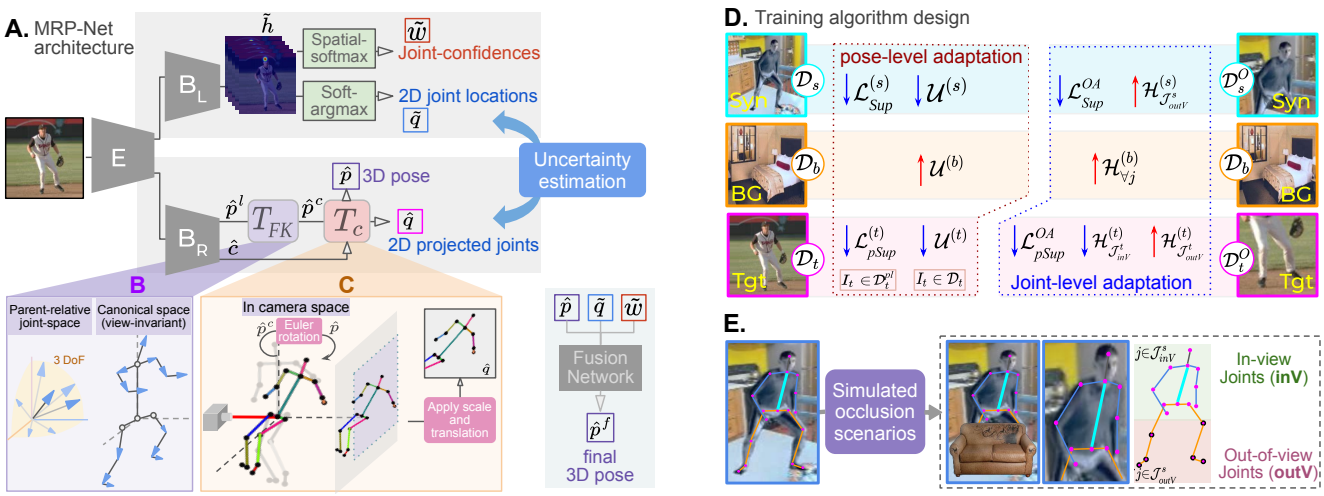

Figure 2. An overview of the proposed framework. **A.** Design configuration for output representations of *MRP-Net* architecture. **B.** Details of the Forward-kinematics transformation. **C.** Applying rotation and camera transformations. **D.** An illustration of the datasets and loss terms for the proposed pose-level and joint-level adaptation. **E.** The occlusion simulation to obtain in-view and out-view joint-ids.

canonical pose coordinates $\hat{p}^c$ does not ensure the 3D bone-length constraints. Thus, we obtain the canonical pose via a forward-kinematic transformation where the pose-network regresses local limb vectors of unit magnitudes, $\hat{p}^l \in \mathbb{R}^{J\times 3}$. For each joint j, the limb-vector is defined at a predefined convention of parent-relative joint space. Here, the forward-kinematic transformation T_{FK} builds the canonical pose by recursively traversing over joints in the kinematic tree; while applying pre-fixed bone-length magnitudes along the transformed limb-vector directions (see Fig. 2B). Alongside the local limb-vectors, the pose-network regresses the Euler-rotations alongside the scale and spatial translation parameters ($\hat{c}$: 3 angles, 1 scale, and 2 translation parameters). Following this, scaled orthographic projection T_c outputs the projected image-space joint coordinates $\hat{q} \in \mathbb{R}^{J\times 2}$. This also outputs the camera-space 3D pose $\hat{p} \in \mathbb{R}^{J\times 3}$ as an intermediate representation.

Next, we quantify uncertainty as follows:

a) Quantifying pose-level uncertainty. In literature, ensemble disagreement provides a useful quantitative measure to evaluate the prediction uncertainty [18]. For *MRP-Net*, we propose to rely on the diversity in design configuration between the two representations obtained via B_L and B_R. Thus, we define the *pose-uncertainty* as follows:

$$\mathcal{U}(I) = |\tilde{q} - \hat{q}|; \quad \tilde{q} = B_L \circ E(I), \quad \hat{q} = T \circ B_R \circ E(I) \quad (1)$$

Here, $\circ$ denotes functional composition and $T = T_{FK} \circ T_c$.

b) Quantifying joint-level uncertainty. Among the two representations, joint-level uncertainty can be extracted from localization based spatial map distributions. For each joint prediction, the joint uncertainty associated with a joint j, is realized as the self-entropy of spatial distributions, *i.e.*

$$\mathcal{H}(I, j) = -\sum_{v} \tilde{h}^{(j)}(v) \log \tilde{h}^{(j)}(v) \quad (2)$$

3.2. Pose-level adaptation framework

In unsupervised DA the primary goal is to transfer the task knowledge from a labeled source dataset $\mathcal{D}_s$ (synthetic domain) to an unlabeled target dataset $\mathcal{D}_t$ (real domain).

3.2.1 Preparing pose-uncertainty-aware *MRP-Net*

Let, $\mathcal{L}_h(.)$ and $\mathcal{L}_p(.)$ be the mean squared loss for the heatmap and the 3D pose respectively. The synthetic supervision loss is expressed as;

$$\mathcal{L}_{Sup}^{(s)}(I \in \mathcal{D}_s) = \mathcal{L}_h(\tilde{h}, h_{gt}) + \lambda_1 \mathcal{L}_p(\hat{p}, p_{gt}) + \lambda_2 \mathcal{U}^{(s)} \quad (3)$$

Here, h_{gt} and p_{gt} denote the respective ground-truths (GT) with λ_1 and λ_2 being the balancing hyperparameters. The intended behaviour of *pose-uncertainty* is that it would elicit a high value for $\mathcal{U}^{(s)}$ for unfamiliar inputs while being low for familiar in-domain samples, *i.e.* for $I \in \mathcal{D}_s$. However, training of *MRP-Net* solely on samples from $\mathcal{D}_s$ outputs consistently low *pose-uncertainty* for both in-domain and out-of-domain samples during validation. One has to explicitly update the network parameters to obtain higher uncertainty for the unfamiliar inputs. In view of the human pose estimation task, we resort to a dataset of background images $\mathcal{D}_b$ (*i.e.* images without any person in frame) to approximate an extreme out-of-distribution scenario.

In summary, the *MRP-Net* is trained to minimize $\mathcal{L}_{Sup}^{(s)}$ while simultaneously maximizing the *pose-uncertainty* for backgrounds *i.e.* $\mathcal{U}^{(b)} = \mathcal{U}(I \in \mathcal{D}_b)$.

3.2.2 Adaptation via uncertainty minimization

Next, the *uncertainty-aware* network is exposed to the unlabeled target samples, $I_t \in \mathcal{D}_t$. Analyzing the histogram from Fig. 1 of the *pose-uncertainties* for samples from $\mathcal{D}_s$,

$\mathcal{D}_b$, and $\mathcal{D}_t$ shows that the uncertainties for $\mathcal{D}_t$ spans a wide-range of values with the same for $\mathcal{D}_s$ and $\mathcal{D}_b$ as peaky distributions at opposite extremes. One can relate the uncertainty gap between the samples from $\mathcal{D}_b$ and $\mathcal{D}_t$ as result of the distinction between *data-uncertainty* (or aleatoric uncertainty) and *knowledge-uncertainty* (or epistemic uncertainty), respectively. Here, *data-uncertainty* refers to the irreducible uncertainty in prediction as a result of noisy input, whereas the *knowledge-uncertainty* refers to the reducible uncertainty elicited as an outcome of the discrepancy in input distributions (*i.e*. the synthetic versus real domains).

a) Adaptation. Motivated by the above discussion, we seek to minimize the *pose-uncertainty* for the target samples, *i.e*. $\boldsymbol{\mathcal{U}^{(t)}} = \mathcal{U}(I \in \mathcal{D}_t)$ alongside minimizing $\boldsymbol{\mathcal{L}^{(s)}_{Sup}}$, while simultaneously maximizing $\boldsymbol{\mathcal{U}^{(b)}} = \mathcal{U}(I \in \mathcal{D}_b)$.

b) Self-training on target pseudo-labels. The literature [46, 103] suggests that target supervision on a reliable pseudo-label subset helps to improve the adaptation performance. In classification tasks, the class predictions of the most confident targets are collected as a reliable pseudo-label subset [76]. In the proposed scenario, the reliability of pseudo-label selection based on the *pose-uncertainty* is highly questionable, as it might be an outcome of the enforced uncertainty minimization loss instead of a genuine learning induced behaviour. Thus, in order to move away from such dependency, we utilize an equivariance consistency based pseudo-label selection criteria. This is realized by applying the most diverse spatial transformation *i.e*. *image-flip*. Essentially, the 2D pose predictions of a given image, I_t and the corresponding flipped image, $I'_t = \mathcal{F}_I(I_t)$ are compared after a left-right joint-id swapping operation $\mathcal{F}_q$. Thus, for each I_t we obtain the following predictions $\hat{q}_t, \tilde{q}_t, \mathcal{F}_q(\hat{q}'_t), \mathcal{F}_q(\tilde{q}'_t)$. Finally, the pseudo-label subset $\mathcal{D}^{pl}_t$ is realized by selecting samples that have an equivariance-consistency less than a preset threshold α^{th}_p *i.e*.,

$$\mathcal{D}^{pl}_t = \{I_t : (|\hat{q}_t - \mathcal{F}_q(\tilde{q}'_t)| + |\tilde{q}_t - \mathcal{F}_q(\hat{q}'_t)|) < \alpha^{th}_p\} \quad (4)$$

Next, we minimize the target pseudo-label supervision loss;

$$\boldsymbol{\mathcal{L}^{(t)}_{pSup}}(I \in \mathcal{D}^{pl}_t) = \sum_{j=1}^{J} \tilde{w}^{(j)}(\mathcal{L}^{(j)}_h(\tilde{h}, h^{pl}_{gt}) + \lambda \mathcal{L}^{(j)}_p(\hat{p}, p^{pl}_{gt})) \quad (5)$$

Here, the supervised joint-wise loss is weighted by the normalized joint-confidences to avoid strong supervision on confusing joint predictions. p^{pl}_{gt} and h^{pl}_{gt} denote the estimated pseudo-label GT (*i.e*. prediction average over the equivariance instances). Here, $\mathcal{D}^{pl}_t$ alongside the pseudo-label GTs are updated at regular intervals during training.

3.3. Joint-level adaptation framework

Most of the prior 3D pose estimation approaches expect full-body visibility without external occlusion. However, in real-world deployment, the camera feed may capture human images having external object occlusions or truncations. In such scenarios, an intended behavior of the model would be to estimate a reasonably well joint localization specifically for the *in-view* joints with lower *joint-uncertainty* values, and higher joint-uncertainties for the *out-view* joints.

3.3.1 Preparing joint-uncertainty-aware *MRP-Net*

Similar to *pose-uncertainty*, one must simulate joint-level uncertainties to enable the model to elicit the above discussed behavior. Note that, training on synthetic full-body images (*i.e*. $\mathcal{D}_s$) or the backgrounds (*i.e*. $\mathcal{D}_b$) is not suitable enough as they do not encourage varying *joint-uncertainties* for the same input instance. Thus, we simulate an *occlusion-aware synthetic dataset*, $\boldsymbol{\mathcal{D}^O_s}$ with segregated set of *in-view* and *out-view* image-joint pairs, denoted as $\mathcal{J}^s_{inV}$ and $\mathcal{J}^s_{outV}$ respectively (see Fig. 2D). Broadly, we simulate occlusion of two kinds, viz, a) occlusion by an external object, and b) truncation of the image frame. We apply the following synthetic supervision loss.

$$\boldsymbol{\mathcal{L}^{OA}_{Sup}}(I \in \mathcal{D}^O_s) = \mathbb{1}_{\mathcal{J}^s_{inV}}(\mathcal{L}^{(j)}_h(\tilde{h}, h_{gt}) + \lambda_1 \mathcal{L}^{(j)}_p(\hat{p}, p_{gt})) - \lambda_2 \boldsymbol{\mathcal{H}^{(s)}_{\mathcal{J}^s_{outV}}} \quad (6)$$

Here, $\mathbb{1}$ denotes an indicator function. The last-term aims to maximize the *joint-uncertainties* only for the *out-view* joints. We also maximize *joint-uncertainties* of all the joints for the backgrounds $\mathcal{D}_b$ by maximizing $\boldsymbol{\mathcal{H}^{(b)}_{\forall j}}$.

3.3.2 Adapting to unlabeled target with occlusion

Next, the *joint-uncertainty-aware* model is exposed to samples from the unlabeled target dataset containing images of varied kinds including full-body, truncated, and occluded samples. We denote this dataset as $\boldsymbol{\mathcal{D}^O_t}$ against the *full-body* target $\mathcal{D}_t$. We follow the adaptation process very similar to that of the proposed pose-level adaptation with the *pose-uncertainties* replaced by the *joint-uncertainties*.

We follow the similar equivariance-based pseudo-label selection criteria to pick the suitable (I_t, j) pairs for creating the target pseudo-label subset, $\mathcal{J}^t_{inV}$. Here, $\mathcal{J}^t_{outV}$ denotes the set of the (I_t, j) pairs having *joint-uncertainty* greater than a preset threshold α^{th}_h. Rest of the (I_t, j) pairs can move either towards $\mathcal{J}^t_{inV}$ or $\mathcal{J}^t_{outV}$ over the course of adaptation training and are thus left untouched (no loss imposed).

$$\mathcal{J}^t_{inV} = \{(I_t, j) : \mathcal{H}(I_t, j)(|\tilde{q}^{(j)}_t - \mathcal{F}^{(j)}_q(\hat{q}'_t)|) < \alpha^{th}_q\} \quad (7)$$

The adaptation training involves minimizing $\boldsymbol{\mathcal{H}^{(t)}_{\mathcal{J}^t_{inV}}}$ while maximizing $\boldsymbol{\mathcal{H}^{(t)}_{\mathcal{J}^t_{outV}}}$ (*joint-uncertainties* for $\mathcal{J}^t_{inV}$ and $\mathcal{J}^t_{outV}$ respectively), alongside minimizing the joint-supervision on target pseudo-labels, *i.e*. $\boldsymbol{\mathcal{L}^{OA}_{pSup}}((I, j) \in \mathcal{J}^t_{inV})$

3.4. Inferring the final 3D pose

In the proposed *MRP-Net*, 3D pose can only be inferred through the B_R branch (*i.e*. $\hat{p}$). However, several recent

Table 2. Quantitative results on 3DPW and 3DHP. Numbers and layout taken from [98]. * denotes inference stage (or online) optimization.

Adaptation type	Methods	H3.6M→3DPW		H3.6M→3DHP		H3.6M→SURREAL		H3.6M→HumanEva	
		MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
General Adaptation	DDC [85]	110.4	75.3	115.6	91.5	117.5	80.1	83.8	64.9
	DAN [52]	107.5	73.2	109.5	89.2	114.2	78.4	78.5	62.7
	DANN [19]	106.3	71.1	107.9	88.0	113.6	77.2	76.3	60.8
	Zhang *et al*. [98]	94.7	63.9	99.3	81.5	103.3	69.1	69.2	53.5
	Ours	**91.9**	**62.1**	**96.2**	**78.6**	**99.6**	**67.2**	**66.8**	**51.9**
Test-time Adaptation	ISO [96]*	-	70.8	-	75.8	-	-	-	-
	BOA [77]*	92.1	58.8	-	77.4	-	-	-	-
	*Ours+ISO**	**89.6**	**57.5**	**92.9**	**76.3**	**96.4**	**65.1**	**65.2**	**50.1**

works [58] advocate for a localization-based representation even for the 3D pose estimation by introducing another output for joint-wise depth-localization. Based on the pros and cons of both the modeling configurations, we decide to leverage the best of both worlds by training a fusion network to realize the final 3D pose prediction $\hat{p}^f$. The fusion network takes three inputs; a) 3D pose predictions via B_R (*i.e*. $\hat{p}$), b) 2D pose prediction via B_L (*i.e*. $\tilde{q}$), and c) the joint-confidences $\tilde{w}$. The fully-connected fusion network is trained to minimize a loss that is exactly similar to $\boldsymbol{\mathcal{L}}_{pSup}^{(t)}$ but on samples from both source and target pseudo-label set. Please refer to Supplementary for more details.

4. Experiments

We demonstrate effectiveness of *MRP-Net* (*MRPN*) by evaluating it on a variety of cross-dataset settings.

Implementation details. We use *ResNet-50* [25] (till Res-4f), pre-trained on the *ImageNet*, as the common encoder E. The localization branch B_L comprises of of transposed convolutional layers which progressively increase the spatial resolution to yield 17 heatmaps of size 56×56. The regression branch B_R consists of a series of fully-connected (FC) layers which later bifurcate into two sub-branches to yield camera parameters $\hat{c}$ and local limb vectors $\hat{p}^l$. We trained the framework on a NVIDIA P-100 GPU (16GB), with a batch size of eight. We employ separate Adam optimizers [31] for each loss term. See Suppl. for more details.

Datasets. We use the following datasets.

a) SURREAL (S) synthetic dataset [86] is used both as source and target under different problem settings. Though the dataset encapsulates a wide range of diversity, synthetic-trained model suffers from poor generalization on natural images due to synthetic-to-real domain-shift.

b) Backgrounds. We use background images taken from; LSUN [92], Google Street View [93], Natural Scenes [21], and Campus Scenes [8] to form the dataset, $\mathcal{D}_b$.

c) Human3.6M (H). For a fair evaluation, we use the standard, in-studio Human3.6M (H3.6M) dataset [28] as either source or target domain in different problem settings.

d) Target datasets. 3DPW [87], HumanEva [78], and MPI-INF-3DHP (3DHP) [54] are used as unlabeled target datasets to evaluate the unsupervised adaptation.

Table 3. Quantitative comparison on Human3.6M. Our proposed method outperforms the prior-arts at various supervision levels. * denotes using MPII [2] with 2D pose annotations. + denotes using additional in-the-wild data taken from the Internet. supervision-type on target (H3.6M) is indicated under the *Supervision* column. *Semi-sup (S1)* denotes 3D pose supervision only on subject S1.

Methods	Supervision	PA-MPJPE↓	MPJPE↓
Martinez *et al*. [53]	Full-3D	52.5	67.5
Xu *et al*. [90]	Full-3D	36.2	45.6
Chen *et al*. [12]	Full-3D	32.7	47.3
Mitra *et al*. [57]	Semi-sup (S1)	90.8	120.9
Li *et al*. [50]	Semi-sup (S1)	66.5	88.8
Rhodin *et al*. [70]	Semi-sup (S1)	65.1	-
Kocabas *et al*. [32]	Semi-sup (S1)	60.2	-
Iqbal *et al*. [29]*	Semi-sup (S1)	51.4	62.8
Ours(S→H, Semi)	Semi-sup (S1)	**49.6**	**59.4**
Kundu *et al*. [43]+	No sup.	99.2	-
Ours(S→H)	No sup.	**88.9**	**103.2**

Occlusion simulation. We perform occlusion simulation on both source and target. The simulation process works on the corresponding full-body instances. We paste external objects (like cars, chair, wardrobe, etc.) to simulate occlusions, whereas truncation is simulated by randomly zooming into the top or bottom region of the full-body images.

Evaluation metrics. For a fair comparison, we evaluate our approach on the standard benchmark datasets described above. The standard mean per joint position error metric computed before and after Procrustes Alignment [22] are denoted as MPJPE and PA-MPJPE respectively [28].

4.1. Quantitative analysis

4.1.1. Synthetic to real adaptation. Labeled synthetic SURREAL (S) is used as the source domain, while the unlabeled target instances are taken from Human3.6M (H). Table 3 shows a comparison of *Ours(S→H)* on unsupervised and semi-supervised settings. Here, semi-supervised setting denotes 3D supervision only on subject S1. In spite of the huge domain shift between SURREAL and Human3.6M our adaptation strategy yields *state-of-the-art* performance among the semi-supervised and unsupervised prior-arts.

4.1.2. In-studio to in-the-wild adaptation. We evaluate the performance of *MRPN* in Table 2 on four target

Figure 3. Qualitative analysis. 3D poses shown correspond to the original camera view and another azimuthal view (+30° or -30°). For results in panel **E** and **F** the joints with uncertainty greater than a prefix threshold are highlighted with red-blobs. The model fails on rare poses, complex inter-limb occlusion and heavy background clutter as highlighted by red bases. Refer Suppl. for more results.

domains, i.e., 3DHP, 3DPW, SURREAL, and HumanEva datasets when using Human3.6M as the source domain. Our baseline is trained only on the source domain Human3.6M. A direct transfer on the target domains performs poorly attributed to the vast domain gap induced due to changing pose, appearance, and backgrounds between source and target datasets. Our pose-level adaptation strategy helps *MRPN* improve upon the prior-arts even on in-the-wild 3DPW dataset by a significant margin, thereby validating ours superior generalizability. *Ours+ISO* is the variant which uses ISO [96] for test time optimization.

4.1.3. Adaptation to partial body visibility. Table 4 reports a quantitative analysis to highlight the merits of our design choices against the standard prior-art techniques. Under Joint-level adaptation, the baseline on row-1 shows transfer results on the target before adaptation. Row-2 baseline employs uncertainty maximization on target as well. Finally, row-3 uses target pseudo labels for self-training. *Ours(JU)* outperforms LCR++ [74] even in the absence of 2D/3D supervision on in-the-wild datasets like MPII.

4.1.4. Ablation study. In Table 4, under *pose-level* adaptation, the baseline on row-5 uses *MRPN* architecture while employing an adversarial discriminator based discrepancy minimization on encoder features (DANN). The baseline on row-6 shows transfer results on the target before adaptation. Row-8 baseline employees self-training unlike in row-7 where only the target uncertainty is minimized. Our final model is depicted in row-9 where the pose predictions are obtained via the fusion-network unlike in row-4.

4.2. Qualitative evaluation and limitations

Fig. 3 and Fig. 5 analyze our pose prediction results across variations in pose complexity, occlusion/truncation scenarios and environmental conditions (*i.e.* in-studio and in-the-wild). Fig. 3 presents extensive in-the-wild and partial body visibility scenarios. We also show results on images randomly taken from online sources. *MRP-Net* successfully estimates reasonable 3D poses for most of the occluded and truncated cases. However, our model may fail under certain drastic scenarios such as multi-level body-part occlusion, high background clutter, and rare athletic poses. Fig. 5 gives an insight into how *MRP-Net* adapts to both unoccluded and partial-body visibility scenarios with predictions better than LCR++ [74]. *MRPN(B1)* indicates the occlusion-aware network before the adaptation training. Without adaptation on target samples, the model predicts with high uncertainty. *MRPN(PU)* and *MRPN(JU)*

Table 4. Ablation study. The column headings indicate usage of different loss terms during training. Ablations under *pose-level* adaptation are evaluated on Human3.6M. * denotes inference without the fusion network. Ablations under *joint-level* adaptation are evaluated on truncated/occluded 3DPW test-split, obtained via in-house occlusion simulations. Here, MPJPE is computed only for the true *in-view* joints. B1 and B2 denotes our baselines under joint-level and pose-level adaptations respectively.

Joint-level adaptation on 3DPW					
No.	Method	$\mathcal{L}_{Sup}^{OA} - \mathcal{H}_{\forall j}^{(b)}$	$\mathcal{H}^{(t)}$	$\mathcal{L}_{pSup}^{OA}$	MPJPE↓
1.	*B1(JU; H→3DPW)*	✓	-	-	191.2
2.	*B1(JU; H→3DPW)*	✓	✓	-	130.7
3.	*Ours(JU; H→3DPW)*	✓	✓	✓	**98.0**
4.	LCR-Net++ [74]	H3.6M(3D), MPII(2D)			104.9
Pose-level adaptation on Human3.6M					
No.	Method	$\mathcal{L}_{Sup}^{(s)} - \mathcal{U}^{(b)}$	$\mathcal{U}^{t}$	$\mathcal{L}_{pSup}^{(t)}$	MPJPE↓
5.	*B2(S→H)*+DANN* [20]	only $\mathcal{L}_{Sup}^{(s)}$	Standard DA		116.8
6.	*B2(S→H)**	✓	-	-	122.4
7.	*B2(S→H)**	✓	✓	-	113.4
8.	*Ours(S→H)**	✓	✓	✓	106.3
9.	*Ours(S→H)*	✓	✓	✓	**103.2**

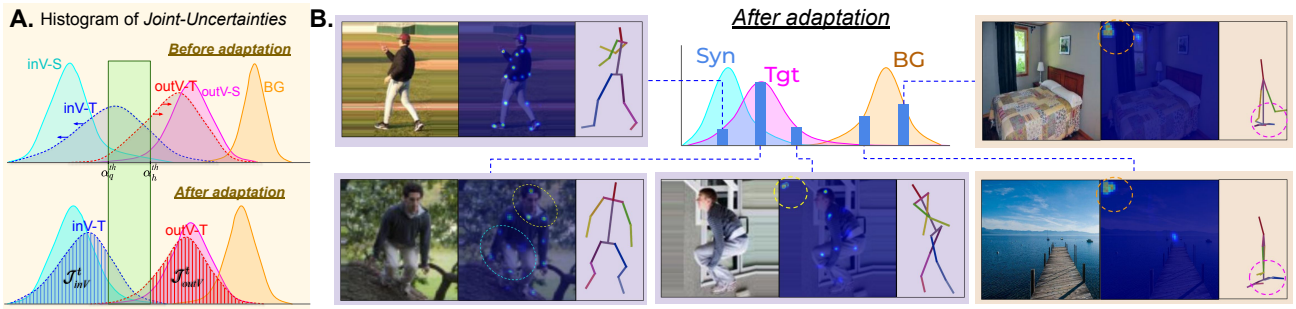

Figure 4. A. Histogram of samples from different domains (source, target, and background) along the joint uncertainty metric. B. Visualizing how the model choose to maximizes uncertainty for background (on right) and and occluded joints (on middle).

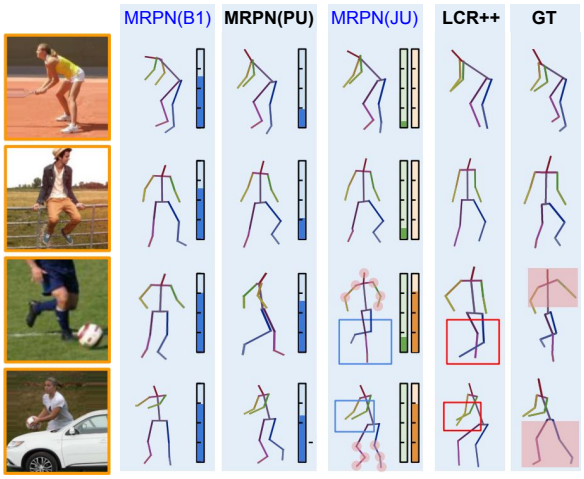

Figure 5. Pose prediction results showing measure of uncertainty values for pose, in-view, and out-view joints. The barometer height indicates high uncertainty. The blue, green and orange barometers indicate the prediction uncertainty for the full-pose, true-in-view joints and true-out-view joints respectively.

indicate the final networks after the *pose-level* and *joint-level* adaptations. *MRPN(PU)* is not tuned to work on occluded/truncated images and thus yields a higher uncertainty for the bottom two images. Whereas, the uncertainty predictions of *MRPN(JU)* for the green and orange barometer yield the expected behaviour. The red-blocks under GT column segregate the true *out-view* joints. The in-view joint predictions of *MRPN(JU)* match with the same under GT.

Societal impacts. We do not foresee a direct negative societal impact from our framework. However, it may be leveraged for human-tracking applications. We urge the readers to make ethical and responsible use of our work.

4.2.1. Model interpretability. We also perform a thorough qualitative study to interpret the behaviour of our network for a wide variety of in-distribution and out-of-distribution samples. In Fig. 4, we analyze how samples from different domains (such as source, target, and backgrounds) are distributed along the uncertainty metrics, *i.e.* the *pose-uncertainty* and *joint-uncertainty*. This gives an insight into how *MRPN* caters to OOD samples as well as the uncertainty associated with partial body visibility. Fig. 4A shows histogram of the predicted *joint-uncertainties* for the true *in-view* and *out-view* joints separately for source (*i.e.* `inV-S` and `outV-S`) and target (*i.e.* `inV-T` and `outV-T`). `BG` denotes the histogram of all *out-view* joints for backgrounds. The shaded regions in the bottom panel depicts $\mathcal{J}^t_{inV}$ and $\mathcal{J}^t_{outV}$ which are segregated using the preset thresholds α^{th}_q and α^{th}_h respectively (edges of the green-box). Our adaptation algorithm succeeds to separate *inV-T* and *outV-T* over the course of adaptation training. Next, Fig. 4B shows a similar analysis for *pose-uncertainties*. We show five different examples sampled from different regions of the histogram-bins. In the right-panel, we show that to maximize *pose-uncertainty* for backgrounds (OOD samples), *MRPN* estimates the 2D landmarks and 3D pose points separated towards opposite diagonal corners. Here, the 2D landmarks are collapsed to the top-left corner whereas the root joint (pelvis) of the model-based 3D predictions are seemed to have collapsed towards the bottom-right corner. In the bottom-panel, for uncertain target instances, we see two peaks in the joint heatmap distributions; one at the top-left corner (OOD-related) and the other near the actual joint location. During adaptation, the OOD-related peak suppresses while the joint-related peak rises to simultaneously reduce the uncertainty while converging towards the true pose outcome. Finally, on the left panel, *joint-level* uncertainty is indicated by the entropy of heatmap distribution.

5. Conclusion

We presented a multi-representation pose network that embraces the pros and cons of both model-free and model-based pose representations to realize a disagreement based pose-uncertainty measure. We develop learning techniques to make the model behave differently for the in-domain and out-of-domain scenarios. Later, the same instigated behaviour is used to devise effective unsupervised adaptation objectives. Formalizing prediction uncertainty in the presence of temporal context remains to be explored in future.

Acknowledgements. This work is supported by Uchhatar Avishkar Yojana (UAY, IISC_010), MoE, Govt. of India.

References

[1] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):44–58, 2005. 3

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 6

[3] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 3

[4] L. Bo and C. Sminchisescu. Structured output-associative regression. In *CVPR*, 2009. 3

[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*, 2016. 3

[6] P. Buehler, A. Zisserman, and M. Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *CVPR*, 2009. 1

[7] M. Burenius, J. Sullivan, and S. Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013. 3

[8] Johannes Burge and Wilson S Geisler. Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences*, 108(40):16849–16854, 2011. 6

[9] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *ICCV*, 2019. 2

[10] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M Rehg. Unsupervised 3D pose estimation with geometric self-supervision. In *CVPR*, 2019. 1

[11] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation. In *CVPR*, 2019. 1

[12] Zerui Chen, Yan Huang, Hongyuan Yu, Bin Xue, Ke Han, Yiru Guo, and Liang Wang. Towards part-aware monocular 3D human pose estimation: An architecture search approach. In *ECCV*, 2020. 6

[13] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3D human pose estimation using spatio-temporal networks with explicit occlusion training. In *AAAI*, 2020. 2

[14] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3D human pose estimation in video. In *ICCV*, 2019. 2

[15] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3D human pose from structure and motion. In *ECCV*, 2018. 1

[16] Rodrigo de Bem, Anurag Arnab, Stuart Golodetz, Michael Sapienza, and Philip Torr. Deep fully-connected part-based models for human pose estimation. In *ACML*, 2018. 2

[17] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3D human pose estimation: motion to the rescue. In *NeurIPS*, 2019. 2

[18] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2, 4

[19] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 1, 6

[20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 7

[21] Wilson S Geisler and Jeffrey S Perry. Statistics for optimal point prediction in natural images. *Journal of Vision*, 11(12):14–14, 2011. 6

[22] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 6

[23] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 3

[24] Nate Hagbi, Oriel Bergig, Jihad El-Sana, and Mark Billinghurst. Shape recognition and pose estimation for mobile augmented reality. *IEEE transactions on visualization and computer graphics*, 17(10):1369–1379, 2010. 1

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[26] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 2

[27] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3D human pose estimation. In *ECCV*, 2018. 3

[28] 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*, 36(7):1325–1339, 2013. 1, 2, 6

[29] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3D human pose learning via multi-view images in the wild. In *CVPR*, 2020. 2, 6

[30] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 3

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

[32] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3D human pose using multi-view geometry. In *CVPR*, 2019. 1, 6

[33] 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. 2

[34] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *ICCV*, 2019. 1

[35] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *ICML*, 2018. 2

[36] Jogendra Nath Kundu, Himanshu Buckchash, Priyanka Mandikal, Rahul MV, Anirudh Jamkhandi, and

R Venkatesh Babu. Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In *WACV*, 2020. 3

[37] Jogendra Nath Kundu, Nishank Lakkakula, and R Venkatesh Babu. UM-Adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *ICCV*, 2019. 1

[38] Jogendra Nath Kundu, Rahul MV, Aditya Ganeshan, and R. Venkatesh Babu. Object pose estimation from monocular image using multi-view keypoint correspondence. In *ECCV Workshops*, 2018. 1

[39] Jogendra Nath Kundu, Rahul MV, Jay Patravali, and R Venkatesh Babu. Unsupervised cross-dataset adaptation via probabilistic amodal 3D human pose completion. In *WACV*, 2020. 2

[40] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul MV, and R. Venkatesh Babu. Appearance consensus driven self-supervised human mesh recovery. In *ECCV*, 2020. 3

[41] Jogendra Nath Kundu, Ambareesh Revanur, Govind V Waghmare, Rahul MV, and R Venkatesh Babu. Unsupervised cross-modal alignment for multi-person 3D pose estimation. In *ECCV*, 2020. 2

[42] Jogendra Nath Kundu, Siddharth Seth, Anirudh Jamkhandi, Pradyumna YM, Varun Jampani, Anirban Chakraborty, and R Venkatesh Babu. Non-local latent relation distillation for self-adaptive 3D human pose estimation. In *NeurIPS*, 2021. 3

[43] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3D human pose estimation via part guided novel image synthesis. In *CVPR*, 2020. 6

[44] Jogendra Nath Kundu, Siddharth Seth, Rahul MV, Rakesh Mugalodi, R Venkatesh Babu, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3D human pose estimation. In *AAAI*, 2020. 3

[45] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 3

[46] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshops*, 2013. 5

[47] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *ICLR*, 2018. 2

[48] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 1

[49] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3D human pose estimation with evolutionary training data. In *CVPR*, 2020. 2

[50] Z. Li, X. Wang, F. Wang, and P. Jiang. On boosting single-frame 3D human pose estimation via monocular videos. In *ICCV*, 2019. 6

[51] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 2

[52] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 6

[53] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 3, 6

[54] 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. 6

[55] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular rgb. In *3DV*, 2018. 2

[56] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2, 3

[57] Rahul Mitra, Nitesh B. Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3D human pose estimation. In *CVPR*, 2020. 6

[58] 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. 2, 6

[59] Francesc Moreno-Noguer. 3D human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017. 1, 3

[60] Jiteng Mu, Weichao Qiu, Gregory D. Hager, and Alan L. Yuille. Learning from synthetic animals. In *CVPR*, 2020. 2

[61] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2, 3

[62] B. X. Nie, P. Wei, and S. Zhu. Monocular 3D human pose estimation by predicting depth on joints. In *ICCV*, 2017. 3

[63] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3DPO: Canonical 3D pose networks for non-rigid structure from motion. In *ICCV*, 2019. 1

[64] 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*, 2019. 3

[65] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018. 3

[66] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 3

[67] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *BMVC*, 2018. 2, 3

[68] Ibrahim Radwan, Abhinav Dhall, and Roland Goecke. Monocular image 3D human pose estimation under self-occlusion. In *ICCV*, 2013. 2

[69] Mugalodi Rakesh, Jogendra Nath Kundu, Varun Jampani, and R Venkatesh Babu. Aligning silhouette topology for self-adaptive 3D human pose recovery. *NeurIPS*, 2021. 1

[70] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *ECCV*, 2018. 2, 6

[71] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3D human pose estimation from multi-view images. In *CVPR*, 2018. 1

[72] Iasonas Kokkinos Riza Alp Guler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 3

[73] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-classification-regression for human pose. In *CVPR*, 2017. 2

[74] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1146–1161, 2019. 2, 7

[75] Rómer Rosales and S. Sclaroff. Learning body pose via specialized maps. In *NeurIPS*, 2001. 3

[76] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, 2017. 5

[77] Guan Shanyan, Xu Jingwei, Wang Yunbo, Ni Bingbing, and Yang Xiaokang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *CVPR*, 2021. 2, 6

[78] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 1, 6

[79] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[80] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 2, 3

[81] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit Agrawal. PoseNet3D: Learning temporally consistent 3D human pose via knowledge distillation. In *3DV*, 2020. 2, 3

[82] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3D human pose estimation fusing video and inertial sensors. In *BMVC*, 2017. 2, 3

[83] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2D-to-3D lifting and image-to-image translation from unpaired supervision. In *ICCV*, 2017. 1

[84] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 1

[85] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *ArXiv*, abs/1412.3474, 2014. 6

[86] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2, 6

[87] 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. 6

[88] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In *CVPR*, 2019. 1

[89] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Incremental adversarial domain adaptation for continually changing environments. In *ICRA*, 2018. 2

[90] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3D human pose estimation. In *CVPR*, 2020. 6

[91] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3D human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 3

[92] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

[93] Amir Roshan Zamir and Mubarak Shah. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1546–1558, 2014. 6

[94] 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*, 2018. 2

[95] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. *NeurIPS*, 2018. 2

[96] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3D human pose estimation. In *NeurIPS*, 2020. 2, 6, 7

[97] Xiheng Zhang, Yongkang Wong, Mohan S. Kankanhalli, and Weidong Geng. Unsupervised domain adaptation for 3D human pose estimation. In *ACMMM*, 2019. 2

[98] Xiheng Zhang, Yongkang Wong, Xiaofei Wu, Juwei Lu, Mohan Kankanhalli, Xiangdong Li, and Weidong Geng. Learning causal representation for training cross-domain pose estimator via generative interventions. In *ICCV*, 2021. 6

[99] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3D human pose regression. In *CVPR*, 2019. 3

[100] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3D human pose estimation in the wild: A weakly-supervised approach. In *ICCV*, 2017. 3

[101] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *ECCV*, 2016. 3

[102] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *CVPR*, 2016. 2

[103] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 5