Kinematic-Structure-Preserved Representation for Unsupervised 3D Human Pose Estimation

Jogendra Nath Kundu,[*] Siddharth Seth,[*] Rahul M V,[*] Mugalodi Rakesh,
R. Venkatesh Babu, Anirban Chakraborty
Indian Institute of Science, Bangalore, India
{jogendrak, siddharthseth, venky, anirban}@iisc.ac.in, rmvenkat@andrew.cmu.edu, rakeshramesha@gmail.com

Abstract

Estimation of 3D human pose from monocular image has gained considerable attention, as a key step to several human-centric applications. However, generalizability of human pose estimation models developed using supervision on large-scale in-studio datasets remains questionable, as these models often perform unsatisfactorily on unseen in-the-wild environments. Though weakly-supervised models have been proposed to address this shortcoming, performance of such models relies on availability of paired supervision on some related task, such as 2D pose or multi-view image pairs. In contrast, we propose a novel kinematic-structure-preserved unsupervised 3D pose estimation framework, which is not restrained by any paired or unpaired weak supervisions. Our pose estimation framework relies on a minimal set of prior knowledge that defines the underlying kinematic 3D structure, such as skeletal joint connectivity information with bone-length ratios in a fixed canonical scale. The proposed model employs three consecutive differentiable transformations namely forward-kinematics, camera-projection and spatial-map transformation. This design not only acts as a suitable bottleneck stimulating effective pose disentanglement, but also yields interpretable latent pose representations avoiding training of an explicit latent embedding to pose mapper. Furthermore, devoid of unstable adversarial setup, we re-utilize the decoder to formalize an energy-based loss, which enables us to learn from in-the-wild videos, beyond laboratory settings. Comprehensive experiments demonstrate our state-of-the-art unsupervised and weakly-supervised pose estimation performance on both Human3.6M and MPI-INF-3DHP datasets. Qualitative results on unseen environments further establish our superior generalization ability.

1 Introduction

Building general intelligent systems, capable of understanding the inherent 3D structure and pose of non-rigid humans from monocular RGB images remains an illusive goal in the vision community. In recent years, researchers aim to solve this problem by leveraging the advances in two key aspects, *i.e.* a) improved architecture design (Newell, Yang, and Deng 2016; Chu et al. 2017) and b) increasing collection of diverse annotated samples to fuel the supervised learning paradigm (Mehta et al. 2017b). However, obtaining 3D pose ground-truth for non-rigid human-bodies is a highly inconvenient process. Available motion capture systems, such as body-worn sensors (IMUs) or multi-camera structure-from-motion (SFM), require careful pre-calibration, and hence usually done in a pre-setup laboratory environment (Ionescu et al. 2013; Zhang et al. 2017). This often restricts diversity in the collected dataset, which in turn hampers generalization of the supervised models trained on such data. For instance, the widely used Human3.6M (Ionescu et al. 2013) dataset captures 3D pose using 4 fixed cameras (*i.e.* only 4 backgrounds scenes), 11 actors (*i.e.* limited apparel variations), and 17 action categories (*i.e.* limited pose diversity). A model trained on this dataset delivers impressive results when tested on samples from the same dataset, but does not generalize to an unknown deployed environment, thereby yielding non-transferability issue.

Table 1: Characteristic comparison of our approach against prior unsupervised and weakly-supervised human 3D pose estimation works, in terms of access to direct (paired) or indirect (unpaired) supervision levels (MV: Multi-View). Note that, in the proposed framework **the latent pose representation itself, is the 3D pose coordinates**, thereby avoiding training of a separate latent to 3D pose mapper (last column).

Methods	Paired sup.			Unpaired sup. (adv. learning)		Sup. for latent to 3D pose mapping
	2D pose	MV pair	Cam. extrin.	2D pose	3D pose	
(Rhodin et al. 2018)	No	Yes	Yes	No	No	Yes
(kocabas et al. 2019)	Yes	Yes	No	No	No	No
(Chen et al. 2019b)	Yes	Yes	No	No	No	Yes
(Wandt et al. 2019)	Yes	No	No	No	Yes	No
(Chen et al. 2019a)	Yes	No	No	Yes	No	No
Ours (unsup.)	No	No	No	No	No	No

To deal with this problem, researchers have started exploring innovative techniques to reduce dependency on annotated real samples. Aiming to enhance appearance diversity on known 3D pose samples (CMU-MoCap), synthetic datasets have been proposed, by compositing a diverse set of human template foregrounds with random backgrounds (Varol et al. 2017). However, models trained on such samples do not generalize to a new motion (e.g. a particular dance form), apparel, or environment much different from the training samples, as a result of large domain

[*] equal contribution

shift. Following a different direction, several recent works propose weakly-supervised approaches (Zhou et al. 2017), where they consider access to a large-scale dataset with paired supervision on some related-tasks other than the task in focus (*i.e.* 3D pose estimation). Particularly, they access multiple cues for weak supervision, such as, a) paired 2D ground-truth, b) unpaired 3D ground-truth (3D pose without the corresponding image), c) multi-view image pair (Rhodin et al. 2018), d) camera parameters in a multi-view setup etc. (see Table 1 for a detailed analysis).

While accessing such weak paired-supervisions, the general approach is to formalize a self-supervised consistency loop, such as 2D→3D→2D (Tung et al. 2017), view-1→3D→view-2 (Kocabas, Karagoz, and Akbas 2019), etc. However, the limitations of domain-shift still persist as a result of using annotated data (*e.g.* 2D ground-truth or multi-view camera extrinsic). To this end, without accessing such paired samples, (Jakab et al. 2019) proposed to leverage unpaired samples to model the natural distribution of the expected representations (*i.e.* 2D or 3D pose) using adversarial learning. Obtaining such samples, however, requires access to a 2D or 3D pose dataset and hence the learning process is still biased towards the action categories presented in that dataset. One can not expect to have access to any of the above discussed paired or unpaired weak supervisory signals for an unknown deployed environment (e.g. frames of a dance-show where the actor is wearing a rare traditional costume). This motivates us to formalize a fully-unsupervised framework for monocular 3D pose estimation, where the pose representation can be adapted to the deployed environment by accessing only the RGB video frames devoid of dependency on any explicit supervisory signal.

Our contributions. We propose a novel unsupervised 3D pose estimation framework, relying on a carefully designed kinematic structure preservation pipeline. Here, we constrain the latent pose embedding to form an interpretable 3D pose representation, thus avoiding the need for an explicit latent to 3D pose mapper. Several recent approaches aim to learn a prior characterizing kinematically plausible 3D human poses using available MoCap datasets (Kundu et al. 2019). In contrast, we plan to utilize minimal kinematic prior information, by adhering to the restrictions to not use any external unpaired supervision. This involves, a) access to the knowledge of hierarchical limb connectivity, b) a vector of allowed bone length ratios, and c) a set of 20 synthetically rendered images with diverse background and pose (*i.e.* a minimal dataset with paired supervision to standardize the model towards the intended 2D or 3D pose conventions). The aforementioned prior information is very minimal in comparison to the pose-conditioned limits formalized by (Akhter et al. 2015) in terms of both dataset size and parameters associated to define the constraints.

In the absence of multi-view or depth information, we infer 3D structure, directly from the video samples, for the unsupervised 3D pose estimation task. One can easily segment moving objects from a video, in absence of any background (BG) motion. However, this is only applicable to in-studio static camera feeds. Aiming to work on in-the-wild YouTube videos , we formalize separate unsupervised learning schemes for videos with both static and dynamic BG. In absence of background motion, we form pairs of video frames with a rough estimate of the corresponding BG image, following a training scheme to disentangle foreground-apparel and the associated 3D pose. However, in the presence of BG motion, we lack in forming such consistent pairs, and thus devise a novel energy-based loss on the disentangled pose and appearance representations. In summary,

- We formalize a novel collection of three differentiable transformations, which not only acts as a bottleneck stimulating effective pose disentanglement but also yields interpretable latent pose representations avoiding training of an explicit latent-to-pose mapper.
- The proposed energy-based loss, not only enables us to learn from in-the-wild videos, but also improves generalizability of the model as a result of training on diverse scenarios, without ignoring any individual image sample.
- We demonstrate *state-of-the-art* unsupervised and weakly-supervised 3D pose estimation performance on both Human3.6M and MPI-INF-3DHP datasets.

2 Related Works

3D human pose estimation. There is a plethora of fully-supervised 3D pose estimations works (Fang et al. 2018; Mehta et al. 2017a; 2017b), where the performance is benchmarked on the same dataset, which is used for training. Such approaches do not generalize on minimal domain shifts beyond the laboratory environment. In absence of large-scale diverse outdoor datasets with 3D pose annotations, datasets with 2D pose annotations are used as a weak supervisory signal for transfer learning using various 2D to 3D lifting techniques (Tung et al. 2017; Chen et al. 2017; Ramakrishna et al. 2012). However, these approaches still rely on availability of 2D pose annotations. Avoiding this, (Kocabas et al. 2019; Rhodin et al. 2018) proposed to use multi-view correspondence acquired by synchronized cameras. But in such approaches (Rhodin et al. 2018), the latent pose representation remains un-interpretable and abstract, thereby requiring a substantially large amount of 3D supervision to explicitly train a *latent-to-pose mapping* mapper. We avoid training of such explicit mapping, by casting the latent representation, itself as 3D pose coordinates. This is realized as a result of formalizing the geometry-aware bottleneck.

Geometry-aware representations. To capture intrinsic structure of objects, the general approach is to disentangle individual factors of variations, such as appearance, camera viewpoint and other pose related cues, by leveraging inter-instance correspondence. In literature, we find unsupervised land-mark detection techniques (Zhang et al. 2018), that aim to utilize a relative transformation between a pair of instances of the same object, targeting the 2D pose estimation task. To obtain such pairs, these approaches rely on either of the following two directions, viz. a) frames from a video with an acceptable time-difference (Jakab et al. 2018), or b) synthetically simulated 2D transformations (Rocco, Arandjelovic, and Sivic 2017). However, such techniques fail to capture the 3D structure of the object in the absence of multi-view information. The problem becomes more challenging for deformable 3D skeletal structures as found in diverse human poses. Recently (Jakab et al. 2018) proposed an unsupervised 2D landmark estimation method to disentangle pose from appearance using a conditional image generation framework. However, the predicted 2D landmarks do not match with the standard human pose key-points, hence are highly un-interpretable with some landmarks even lying on the background. Such outputs can not be used for a consequent task requiring a structurally consistent 2D pose input.

Formalizing structural constraints in 2D is highly ill-posed, considering images as projections of the actual 3D world. Acknowledging this, we plan to estimate 3D pose separately with camera parameters followed by a camera-projection to obtain the 2D landmarks. As a result of this inverse-graphics formalization, we have the liberty to impose structural constraints directly on the 3D skeletal representation, where the bone-length and other kinematic constraints can be imposed seamlessly using consistent rules as compared to the corresponding 2D representation. A careful realization of 3D structural constraints not only helps us to obtain interpretable 2D landmarks but also reduces the inherent uncertainty associated with the process of lifting a monocular 2D image to its 3D pose (Chen et al. 2019a), in absence of any additional supervision such as multi-view or depth cues.

3 Approach

Our aim is to learn a mapping function, that can map an RGB image of human to its 3D pose by accessing minimal kinematic prior information. Motivated by (Rhodin et al. 2018), we plan to cast it as an unsupervised disentanglement of three different factors *i.e.*, a) foreground (FG) appearance, b) background (BG) appearance, and c) kinematic pose. However, unlike (Rhodin et al. 2018) in absence of multi-view pairs, we have access to simple monocular video streams of human actions consisting of both static and dynamic BG.

3.1 Architecture

As shown in Fig. 1A, we employ two encoder networks each with a different architecture, E_P and E_A to extract the local-kinematic parameters v_k (see below) and FG-appearance, f_a respectively from a given RGB image. Additionally, E_P also outputs 6 camera parameters, denoted by c, to obtain coordinates of the camera-projected 2D landmarks, p_{2D}.

One of the major challenges in learning factorized representations (Denton and others 2017) is to realize purity among the representations. More concretely, the appearance representation should not embed any pose related information and vice-versa. To achieve this, we enforce a bottleneck on the pose representation by imposing kinematic-structure based constraints (in 3D) followed by an inverse-graphics formalization for 3D to 2D re-projection. This introduces three pre-defined transformations *i.e.*, a) Forward kinematic transformation, $\mathcal{T}_{fk}$ and b) Camera projection transformation $\mathcal{T}_c$, and c) Spatial-map transformation $\mathcal{T}_m$.

a) Forward kinematic transformation, $\mathcal{T}_{fk}$ Most of the prior 3D pose estimation approaches (Chen et al. 2019a; Rhodin et al. 2018) aim to either directly regress joint locations in 3D or depth associated with the available 2D landmarks. Such approaches do not guarantee validity of the kinematic structure, thus requiring additional loss terms in the optimization pipeline to explicitly impose kinematic constraints such as bone-length and limb-connectivity information (Habibie et al. 2019). In contrast, we formalize a view-invariant local-kinematic representation of the 3D skeleton based on the knowledge of skeleton joint connectivity. We define a canonical rule (see Fig. 1B), by fixing the neck and pelvis joint (along z-axis, with pelvis at the origin) and restricting the trunk to hip-line (line segment connecting the two hip joints) angle, to rotate only about x-axis on the YZ-plane(*i.e.* 1-DOF) in the canonical coordinate system C (*i.e.* Cartesian system defined at the pelvis as origin). Our network regresses one pelvis to hip-line angle and 13 unit-vectors (all 3-DOF), which are defined at their respective parent-relative local coordinate systems, $L^{Pa(j)}$, where $Pa(j)$ denotes the parent joint of j in the skeletal kinematic tree. Thus, $v_k \in \mathbb{R}^{40}$ (*i.e.* 1+13*3). These predictions are then passed on to the forward-kinematic transformation to obtain the 3D joint coordinates p_{3D} in C, *i.e.* $\mathcal{T}_{fk} : v_k \rightarrow p_{3D}$ where $p_{3D} \in \mathbb{R}^{3J}$, with J being the total number of skeleton joints. First, positions of the 3 root joints, $p_{3D}^{(j)}$ for j as left-hip, right-hip and neck, are obtained using the above defined canonical rule after applying the estimate of the trunk to hip-line angle, $v_k^{(0)}$. Let $len^{(j)}$ store the length of the line-segment (in a fixed canonical unit) connecting a joint j with $Pa(j)$. Then, $p_{3D}^{(j)}$ for rest of the joints is realized using the following recursive equation, $p_{3D}^{(j)} = p_{3D}^{(Pa(j))} + len^{(j)} v_k^{(j)}$. See Fig. 1B (dotted box) for a more clear picture.

b) Camera-projection transformation, $\mathcal{T}_c$ As p_{3D} is designed to be view-invariant, we rely on estimates of the camera extrinsics c (3 angles, each predicted as 2 parameters, the $\sin$ and $\cos$ component), which is used to rotate and translate the camera in the canonical coordinate system C, to obtain 2D landmarks of the skeleton (*i.e.* using the rotation and translation matrices, R_c and T_c respectively). Note that, these 2D landmarks are expected to register with the corresponding joint locations in the input image. Thus, the 2D landmarks are obtained as, $p_{2D}^{(j)} = P(R_c * p_{3D}^{(j)} + T_c)$, where P denotes a fixed perspective camera transformation.

c) Spatial-map transformation, $\mathcal{T}_m$ After obtaining coordinates of the 2D landmarks $p_{2D} \in \mathbb{R}^{2J}$, we aim to effectively aggregate it with the spatial appearance-embedding f_a. Thus, we devise a transformation procedure $\mathcal{T}_m$, to transform the vectorized 2D coordinates into spatial-maps denoted by $f_{2D} \in \mathbb{R}^{H \times W \times Ch}$, which are of consistent resolution to f_a, *i.e.* $\mathcal{T}_m : p_{2D} \rightarrow f_{2D}$. To effectively encode both joint locations and their connectivity information, we propose to generate two sets of spatial maps namely, a) heat-map, f_{hm} and b) affinity-map, f_{am} (*i.e.*, $f_{2D} : (f_{hm}, f_{am})$). Note that, the transformations to obtain these spatial maps must be fully differentiable to allow the disentaglement of pose using the cross-pose image-reconstruction loss, computed at the decoder output (discussed in Sec. 3.3a). Keeping

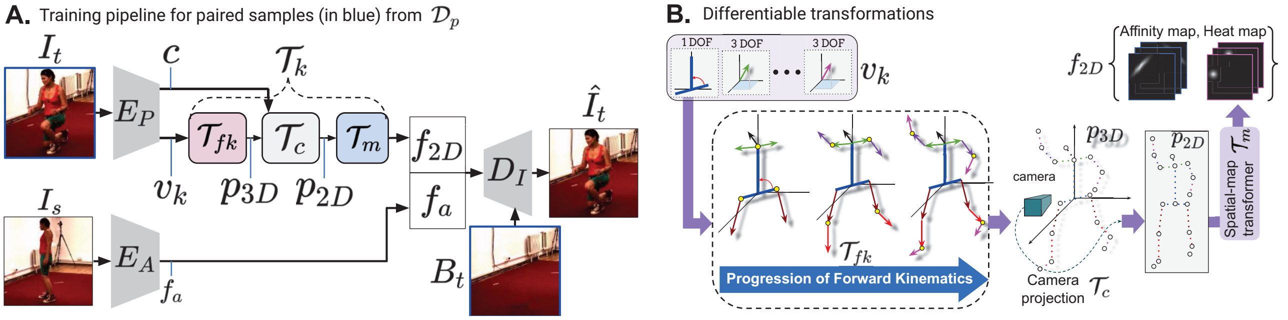

Figure 1: A. Illustration of the proposed framework indicating output notation of individual modules. **B.** An overview of the three differentiable transformations, with step-wise progression of forward kinematics using local-kinematic parameters, v_k.

this in mind, we implement a novel computational pipeline by formalizing translated and rotated Gaussians to represent both joint positions (*i.e.* f_{hm}) and skeleton-limb connectivity (*i.e.* f_{am}). We use a constant variance σ along both spatial directions to realize the heat-maps for each joint j, as $f^{(j)}_{hm}(u) = \exp(-0.5||u - p^{(j)}_{2d}||^2/\sigma^2)$, where $u : [u_x, u_y]$ denotes the spatial-index in a $H \times W$ lattice (see Fig. 2A).

We formalize the following steps to obtain the affinity maps based on the connectivity of joints in the skeletal kinematic tree (see Fig. 2A). For each limb (line-segment), l with endpoints $p^{l(j_1)}_{2D}$ and $p^{l(j_2)}_{2D}$, we first compute location of its mid-point, $\mu^{(l)} : [\mu^{(l)}_x, \mu^{(l)}_y]$ and slope $\theta^{(l)}$. Following this, we perform an affine transformation to obtain, $u' = R_{\theta^{(l)}} * (u - \mu^{(l)})$, where $R_{\theta^{(l)}}$ is the 2D rotation matrix. Let, $\sigma^{(l)}_x$ and $\sigma^{(l)}_y$ denote variance of a Gaussian along both spatial directions representing the limb l. We fix $\sigma^{(l)}_y$ from prior knowledge of the limb width, whereas, $\sigma^{(l)}_x$ is computed as $\alpha * len(l)$ in the 2D euclidean space. Finally, the affinity map is obtained as,

$$f^{(l)}_{am}(u) = \exp(-0.5||u'_x/\sigma^{(l)}_x||^2 - 0.5||u'_y/\sigma^{(l)}_y||^2)$$

$\mathcal{T}_{fk}$, $\mathcal{T}_c$ and $\mathcal{T}_m$ (collectively denoted as $\mathcal{T}_k$) are designed using perfectly differentiable operations, thus allowing back-propagation of gradients from the loss functions defined at the decoder output. As shown in Fig. 1A, the decoder takes in a tuple of spatial-pose-map representation and appearance (f_{2D} and f_a respectively, concatenated along the channel dimension) to reconstruct an RGB image. To effectively disentangle BG information in f_a, we fuse the background image B_t towards the end of decoder architecture, inline with (Rhodin et al. 2018).

3.2 Access to minimal prior knowledge

One of the key objectives of this work is to solve the unsupervised pose estimation problem with minimal access to prior knowledge whose acquisition often requires manual annotation or a data collection setup such as CMU-MoCap. Adhering to this we restrict the proposed framework from accessing any paired or unpaired data samples as shown in Table 1. Here, we list the specific prior information that has been considered in the proposed framework,

- Kinematic skeletal structure (*i.e.* the joint connectivity information) with bone-length ratios in a fixed canonical scale. Note that, we do not consider access to the kinematic angle limits for the limb joints, as such angles are highly pose dependent particularly for diverse human skeleton structures (Akhter and Black 2015).
- A set of 20 synthetically rendered SMPL models with diverse 3D poses and FG appearance (Varol et al. 2017). We have direct paired supervision loss (denoted by $\mathcal{L}_{prior}$) on these samples to standardize the model towards the intended 2D or 3D pose conventions.

3.3 Unsupervised training procedure

In contrast to (Jakab et al. 2018), we aim to disentangle foreground (FG) and background (BG) appearances, along with the disentanglement of pose. In a generalized setup, we also aim to learn from in-the-wild YouTube videos in contrast to in-studio datasets, avoiding dataset-bias.

Separating paired and unpaired samples. For an efficient disentanglement, we aim to form image tuples of the form (I_s, I_t, B_t). Here, I_s and I_t are video frames, which have identical FG-appearance with a nonidentical *kinematic-pose* (pairs formed between frames beyond a certain time-difference). As each video-clip captures action of an individual in a certain apparel, *FG-appearance* remains identical among frames from the same video. Here, B_t denotes an estimate of BG image without the human subject corresponding to the image I_t, which is obtained as the median of pixel intensities across a time-window including the frame I_t. However, such an estimate of B_t is possible only for scenarios with no camera movement beyond a certain time window to capture enough background evidence (*i.e.* static background with a moving human subject).

Given an in-the-wild dataset of videos, we classify temporal clips of a certain duration (>5 seconds) into two groups based on the amount of BG motion in that clip. This is obtained by measuring the pixel-wise L2 loss among the frames in a clip, considering human action covers only 10-20% of pixels in the full video frame. Following this, we realize two disjoint datasets denoted by $\mathcal{D}_p = \{(I^{(i)}_s, I^{(i)}_t, B^{(i)}_t)\}^N_{i=1}$ and $\mathcal{D}_{unp} = \{(I^{(k)}_s, I^{(k)}_t)\}^M_{k=1}$ as

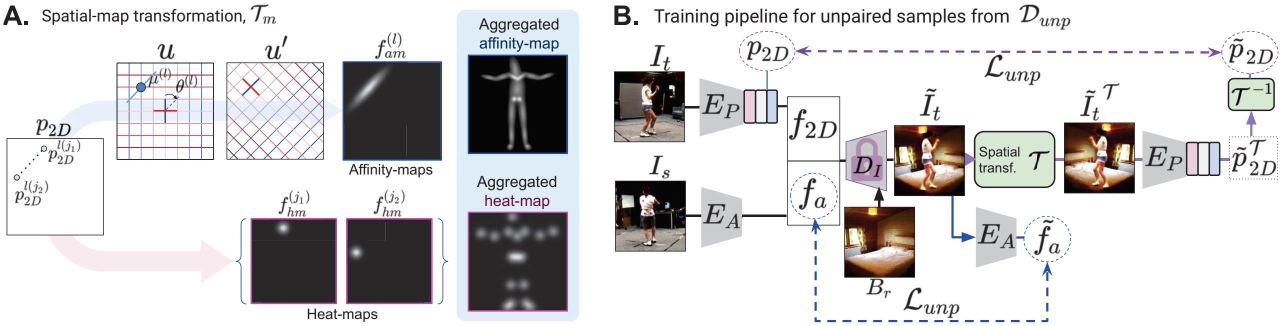

Figure 2: A. Illustration of the steps to obtain the spatial heat-map and affinity-map from the projected 2D coordinates. **B.** An overview of the proposed data-flow pipeline enabling energy-based loss formalization targeting unpaired samples from $\mathcal{D}_{unp}$.

sets of tuples with extractable BG pair (paired) and unextractable BG pair (unpaired), respectively.

a) Training objective for paired samples, $\mathcal{D}_p$ As shown in Fig. 1A, given a source and target image (*i.e.* I_s and I_t), we aim to transfer the pose of I_t (*i.e.* f_{2D}) to the FG-appearance extracted from I_s (*i.e.* f_a) and background from B_t to reconstruct $\hat{I}_t$. Here, the FG and BG appearance information can not leak through pose representation because of the low dimensional bottleneck *i.e.* $p_{2D} \in \mathbb{R}^{2J}$. Moreover, consecutive predefined matrix and spatial-transformation operations further restrict the framework from leaking appearance information through the pose branch even as low-magnitude signals. Note that, the BG of I_s may not register with the BG of I_t, when the person moves in the 3D world (even in a fixed camera scenario) as these images are outputs of an off-the shelf person-detector. As a result of this BG disparity and explicit presence of the clean spatially-registered background B_t, D_I catches the BG information directly from B_t, thereby forcing f_a to solely model FG-appearance from the apparel-consistent source, I_s. Besides this, we also expect to maintain perceptual consistency between I_t and $\hat{I}_t$ through the encoder networks, keeping in mind the later energy-based formalization (next section). Thus, all the network parameters are optimized for the paired samples using the following loss function, $\mathcal{L}_P = |I_t - \hat{I}_t| + \lambda_1|p_{2D} - \hat{p}_{2D}| + \lambda_2|f_a - \hat{f}_a|$. Here, $\hat{p}_{2D} = \mathcal{T}_k \circ E_P(\hat{I}_t)$ and $\hat{f}_a = E_A(\hat{I}_t)$.

b) Training objective for unpaired samples, $\mathcal{D}_{unp}$ Although, we find a good amount of YouTube videos where human motion (e.g. dance videos) is captured on a tripod mounted static camera, such videos are mostly limited to indoor environments. However, a diverse set of human actions are captured in outdoor settings (e.g. sports related activities), which usually involves camera motion or dynamic BG. Aiming to learn a general pose representation, instead of ignoring the frames from video-clips with dynamic BG, we plan to formalize a novel direction to adapt the parameters of E_P and E_A even for such diverse scenarios.

We hypothesize that the decoder D_I expects the pose and FG-appearance representation in a particular form, satisfying the corresponding input distributions, $P(f_{2D})$ and $P(f_a)$. Here, a reliable estimate of $P(f_{2D})$ and $P(f_a)$ can be achieved solely on samples from $\mathcal{D}_p$ in presence of paired supervision, avoiding *mode-collapse*. More concretely, the parameters of D_I should not be optimized on samples from $\mathcal{D}_{unp}$ (as shown in Fig. 2B with a lock sign). Following this, one can treat D_I analogous to a *critic*, which outputs a reliable prediction (an image of human with pose from I_t, FG-appearance from I_s and BG from B_t) only when its inputs f_{2D} and f_a satisfy the expected distributions- $P(f_{2D})$ and $P(f_a)$ respectively. We plan to leverage this analogy to effectively use the frozen D_I network as an energy function to realize simultaneous adaptation of E_P and E_A for the unpaired samples from $\mathcal{D}_{unp}$.

We denote B_r to represent a random background image. As shown in Fig. 2B, here $\tilde{I}_t = D_I(f_{2D}, f_a, B_r)$, in absence of access to a paired image to enforce a direct pixel-wise loss. Thus, the parameters of E_P and E_A are optimized for the unpaired samples using the following loss function, $\mathcal{L}_{UNP} = |p_{2D} - \tilde{p}_{2D}| + \lambda_2|f_a - \tilde{f}_a|$, where $\tilde{p}_{2D} = \mathcal{T}^{-1} \circ \mathcal{T}_k \circ E_P \circ \mathcal{T}(\tilde{I}_t)$ and $\tilde{f}_a = E_A(\tilde{I}_t)$. Here, $\mathcal{T}$ and $\mathcal{T}^{-1}$ represents a differentiable spatial transformation (such as image flip or in-plane rotation) and its inverse, respectively. We employ this to maintain a consistent representation across spatial-transformations. Note that, for the flip-operation of p_{2D}, we also exchange the indices of the joints associated with the left side to right and vice-versa.

We train on three different loss functions, viz. $\mathcal{L}_{prior}$, $\mathcal{L}_P$, and $\mathcal{L}_{UNP}$ at separate iterations, each with different optimizer. Here, $\mathcal{L}_{prior}$ denotes the supervised loss directly on p_{3D} and p_{2D} for the synthetically rendered images on randomly selected backgrounds, as discussed before.

4 Experiments

In this section, we describe experimental details followed by a thorough analysis of the framework for bench-marking on two widely used datasets, Human3.6M and MPI-INF-3DHP.

We use Resnet-50 (till *res4f*) with ImageNet-pretrained parameters as the base pose encoder E_P, whereas the appearance encoder is designed separately using 10 Convolutions. E_P later divides into two parallel branches of fully-connected layers dedicated for v_k and c respectively. We use $J = 17$ for all our experiments as shown in Fig. 1. The

Table 2: Results on Human3.6M following the standard protocol-II setup. Here, *Sup.* (2nd column) denotes the amount of supervision accessed by the respective approaches. Accordingly, the table is divided into 4 row-groups, **a)** row 1-5 use full 3D pose sup., **b)** row 6-10 use full 2D pose as weak sup. **c)** row 11-12: unsupervised approaches, and **d)** row 13: *Ours(semi-sup.)*. We outperform prior approaches in both weakly supervised and unsupervised settings (highlighted as boldface).

Protocol-II	Sup.	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.(↓)
(Akhter et al. 2015)	Full-3D	199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3	160.7	173.7	177.8	181.9	198.6	176.2	192.7	181.1
(Zhou et al. 2016)	Full-3D	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1	137.5	106.0	102.2	110.4	106.5	115.2	106.7
(Bogo et al. 2016)	Full-3D	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	79.7	86.8	87.7	82.3
(Moreno et al. 2017)	Full-3D	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	78.0	71.5	73.2	74.0
(Martinez et al. 2017)	Full-3D	44.8	52.0	44.4	50.5	61.7	59.4	45.1	41.9	66.3	77.6	54.0	58.8	35.9	49.0	40.7	52.1
(Wu et al. 2016)	Full-2D	78.6	90.8	92.5	89.4	108.9	112.4	77.1	106.7	127.4	139.0	103.4	91.4	79.1	-	-	98.4
(Tung et al. 2017)	Full-2D	77.6	91.4	89.9	88.0	107.3	110.1	75.9	107.5	124.2	137.8	102.2	90.3	78.6	-	-	97.2
(Chen et al. 2019a)	Full-2D	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	68.0
(Wandt et al. 2019)	Full-2D	53.0	58.3	59.6	66.5	72.8	71.0	56.7	69.6	78.3	95.2	66.6	58.5	63.2	57.5	49.9	65.1
Ours (weakly-sup.)	Full-2D	56.0	53.2	56.3	63.6	74.1	77.5	53.4	67.9	75.8	90.8	64.2	56.9	61.4	56.3	49.7	**63.8**
(Rhodin et al. 2018)	Multi-view	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	98.2
Ours (unsup.)	**No sup.**	80.2	81.3	86.0	86.7	94.1	83.4	87.5	84.2	101.2	110.9	86.0	87.8	86.9	94.3	90.9	**89.4**
Ours (semi-sup.)	5%-3D	46.6	54.5	50.1	46.4	81.3	42.4	41.1	56.4	86.7	82.9	49.0	47.7	64.1	48.2	44.3	**56.1**

Table 3: Results for the MPI-INF-3DHP dataset. Here, *Trainset* (2nd column) denotes access to 3DHP trainset images before evaluation. *Sup.* (3rd column) denotes supervision level on 3DHP image-pose pairs. 4 row-groups, **a)** row 1-2: Fully supervised, **b)** row 3-7: Weakly supervised, **c)** row 8-10: Unsupervised, **d)** row 11: Semi-supervised.

No.	Method	Trainset	Sup.	PCK (↑)	AUC (↑)	MPJPE (↓)
1.	(Mehta et al. 2017c)	+3DHP	Full-3D	76.6	40.4	124.7
2.	(Rogez et al. 2017)	+3DHP	Full-3D	59.6	27.6	158.4
3.	(Zhou et al. 2017)	+3DHP	Full-2D	69.2	32.5	137.1
4.	(Kanazawa et al. 2018)	+3DHP	Full-2D	77.1	40.7	113.2
5.	(Yang et al. 2018)	+3DHP	Full-2D	69.0	32.0	-
6.	(Chen et al. 2019a)	+3DHP	Full-2D	71.7	36.3	-
7.	Ours (weakly-sup.)	+3DHP	Full-2D	**80.2**	**44.8**	**97.1**
8.	(Chen et al. 2019a)	-3DHP	-	64.3	31.6	-
9.	Ours (unsup.)	-3DHP	-	76.5	39.8	115.3
10.	Ours (unsup.)	+3DHP	**No sup.**	**79.2**	**43.4**	**99.2**
11.	Ours (semi-sup.)	+3DHP	5%-3D	81.9	52.6	89.8

channel-wise aggregation of f_{am} (16-channels) and f_{hm} (17-channels) is passed through two convolutional layers to obtain f_{2D} (128-maps), which is then concatenated with f_a (512-maps) to form the input for D_I (each with 14×14 spatial dimension). Our experiments use different AdaGrad optimizers (learning rate: 0.001) for each individual loss components in alternate training iterations, thereby avoiding any hyper-parameter tuning. We perform several augmentations (color jittering, mirroring, and in-plane rotation) of the 20 synthetic samples, which are used to provide a direct supervised loss at the intermediate pose representations.

Datasets. The *base-model* is trained on a mixture of two datasets, *i.e.* Human3.6M and an in-house collection of YouTube videos (also referred as YTube). In contrast to the in-studio H3.6M dataset, YTube contains human subjects in diverse apparel and BG scenes performing varied forms of motion (usually dance forms such as western, modern, contemporary etc.). Note that all samples from H3.6M contribute to the paired dataset $\mathcal{D}_p$, whereas ∼40% samples in YTube contributed to $\mathcal{D}_p$ and rest to $\mathcal{D}_{unp}$ based on the associated BG motion criteria. However, as we do not have ground-truth 3D pose for the samples from YTube (in-the-wild dataset), we use MPI-INF-3DHP (also referred as 3DHP) to quantitatively benchmark generalization of the proposed pose estimation framework.

a) Evaluation on Human3.6M. We evaluate our framework on protocol-II, after performing scaling and rigid alignment of the poses inline with the prior arts (Chen et al. 2019a; Rhodin et al. 2018). We train three different variants of the proposed framework *i.e.* a) *Ours(unsup.)*, b) *Ours(semi-sup.)*, and c) *Ours(weakly-sup.)* as reported in Table 2. After training the *base-model* on the mixed YTube+H3.6M dataset, we finetune it on the static H3.6M dataset by employing $\mathcal{L}_{prior}$ and $\mathcal{L}_p$ (without using any multi-view or pose supervision) and denote this model as *Ours(unsup.)*. This model is further trained with full supervision on the 2D pose landmarks simultaneously with $\mathcal{L}_{prior}$ and $\mathcal{L}_p$ to obtain *Ours(weakly-sup.)*. Finally, we also train *Ours(unsup.)* with supervision on 5% 3D of the entire trainset simultaneously with $\mathcal{L}_{prior}$ and $\mathcal{L}_p$ (to avoid over-fitting) and denote it as *Ours(semi-sup.)*. As shown in Table 2, *Ours(unsup.)* clearly outperforms the prior-art (Rhodin et al. 2018) with a significant margin (89.4 vs. 98.2) even without leveraging multi-view supervision. Moreover, *Ours(weakly-sup.)* demonstrates state-of-the-art performance against prior weakly supervised approaches.

b) Evaluation on MPI-INF-3DHP. We aim to realize a higher level of generalization as a consequence of leveraging rich kinematic prior information. The proposed framework outputs 3D pose, which is bounded by the kinematic plausibility constraints even for unseen apparel, BG and action categories. This characteristic is clearly observed while evaluating performance of our framework on unseen 3DHP dataset. We take *Ours(weakly-sup.)* model trained on YTube+H3.6M dataset to obtain 3D pose predictions on unseen 3DHP testset (9th row in Table 3). We clearly outperform the prior work (Chen et al. 2019a) by a significant margin in a fully-unseen setting (8th and 9th row with -3DHP in Table 3). Furthermore, our weakly supervised model (with 100% 2D pose supervision) achieves state-of-the-art performance against prior approaches at equal supervision levels.

c) Ablation study. In the proposed framework, our major contribution is attributed to the design of differentiable transformations and an innovative way to facilitate the usage of unpaired samples even in presence of BG motion. Though effectiveness of camera-projection has been studied

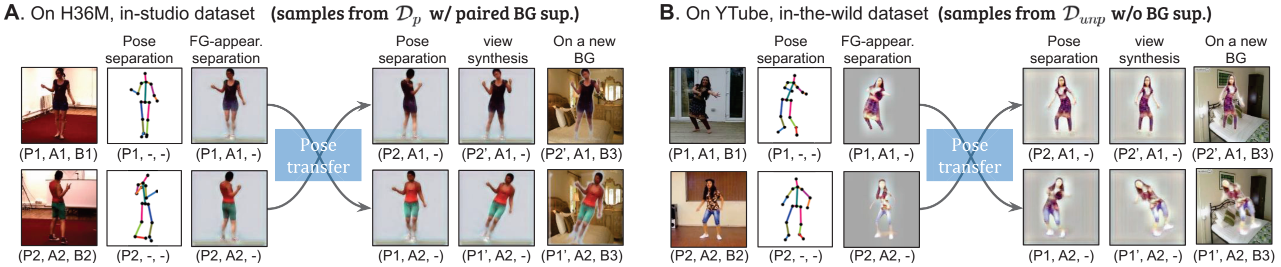

Figure 3: Qualitative results, showing disentanglement of Pose (ID'd as P1 and P2), FG (ID'd as A1 and A2) and BG (ID'd as B1, B2, and B3). Images in first column (of each panel) define the IDs which are later used for novel image synthesis.

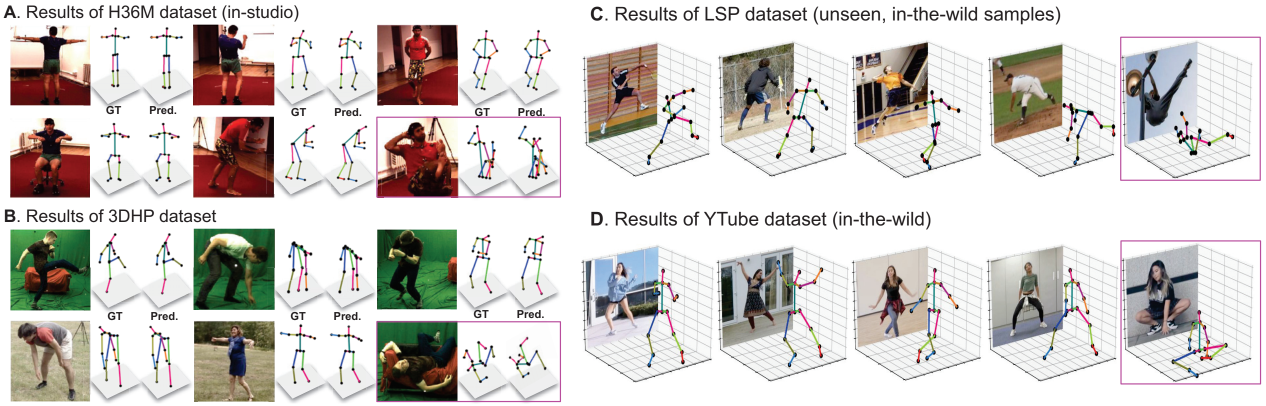
A. Results of H36M dataset (in-studio)
GT
Pred.
GT
Pred.
GT
Pred.
B. Results of 3DHP dataset
GT
Pred.
GT
Pred.
GT
Pred.
C. Results of LSP dataset (unseen, in-the-wild samples)
D. Results of YTube dataset (in-the-wild)

Figure 4: Qualitative results. Note that, results on LSP is obtained in an unseen setting (*i.e.* not even unpaired unsup. training). The pink box highlights the failure cases, specifically in presence of self-occlusion as a result of joint-position ambiguity.

Table 4: Results on ablations of the proposed framework. It clearly highlights importance of $\mathcal{T}_{fk}$, $\mathcal{T}_m$, and use of $\mathcal{D}_{unp}$ in the unsupervised training pipeline. Notice the improvement in 3DPCK on the unseen 3DHP testset as a result of incorporating $\mathcal{D}_{unp}$ in the unsupervised training pipeline.

Method (unsup.)	Training set YTube+H3.6M	MPJPE on H36M	3DPCK on MPI-3DHP
Ours w/o $\mathcal{T}_{fk}$	$\mathcal{D}_p$	134.8	47.9
Ours w/o $\mathcal{T}_m$	$\mathcal{D}_p$	101.8	61.7
Ours(unsup.)	$\mathcal{D}_p$	91.1	66.3
Ours(unsup.)	$\mathcal{D}_p \cup \mathcal{D}_{unp}$	**89.4**	**71.2**

in certain prior works (Chen et al. 2019a), use of forward-kinematic transformation $\mathcal{T}_{fk}$ and affinity map in the spatial-map transformation $\mathcal{T}_m$ is employed for the first time in such a learning framework. Therefore, we evaluate importance of both $\mathcal{T}_{fk}$ and $\mathcal{T}_m$ by separately bypassing these modules through neural network transformations. Results in Table 4 clearly highlight effectiveness of these carefully designed transformations for the unsupervised 3D pose estimation task.

d) Qualitative results. Fig. 3 depicts qualitative results derived from *Ours(unsup.)* on in-studio H3.6M and in-the-wild YTube dataset. It highlights effectiveness of unsupervised disentanglement through separation or cross-transfer of apparel, pose, camera-view and BG, for novel image synthesis. Though, our focus is to disentangle 3D pose information, separation of apparel and pose transfer is achieved as a byproduct of the proposed learning framework. In Fig. 4 we show results on the 3D pose estimation task obtained from *Ours(weakly-sup.)* model. Though we train our model on H3.6M, 3DHP and YTube datasets, results on LSP dataset (Johnson and Everingham 2010) are obtained without training on the corresponding train-set, *i.e.* in a fully-unseen setting. Reliable pose estimation on such diverse unseen images highlights generalization of the learned representations thereby overcoming the problem of dataset-bias.

5 Conclusion

We present an unsupervised 3D human pose estimation framework, which relies on a minimal set of prior knowledge regarding the underlying kinematic 3D structure. The proposed local-kinematic model indirectly endorses a kinematic plausibility bound on the predicted poses, thereby limiting the model from delivering implausible pose outcomes. Furthermore, our framework is capable of leveraging knowledge from video frames even in presence of background motion, thus yielding superior generalization to unseen environments. In future, we would like to extend such frameworks for predicting 3D mesh, by characterizing the prior knowledge on human shape, alongside pose and appearance.

Acknowledgements. This work was supported by a Wipro PhD Fellowship (Jogendra) and in part by DST, Govt. of India (DST/INT/UK/P-179/2017).

References

Akhter, I., and Black, M. J. 2015. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proc. CVPR*.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proc. ECCV*.

Chen, C.-H., and Ramanan, D. 2017. 3d human pose estimation= 2d pose estimation+ matching. In *Proc. CVPR*.

Chen, C.-H.; Tyagi, A.; Agrawal, A.; Drover, D.; MV, R.; Stojanov, S.; and Rehg, J. M. 2019a. Unsupervised 3d pose estimation with geometric self-supervision. In *Proc. CVPR*.

Chen, X.; Lin, K.-Y.; Liu, W.; Qian, C.; and Lin, L. 2019b. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *Proc. CVPR*.

Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A. L.; and Wang, X. 2017. Multi-context attention for human pose estimation. In *Proc. CVPR*.

Denton, E. L., et al. 2017. Unsupervised learning of disentangled representations from video. In *Proc. NeurIPS*.

Fang, H.-S.; Xu, Y.; Wang, W.; Liu, X.; and Zhu, S.-C. 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proc. AAAI*.

Habibie, I.; Xu, W.; Mehta, D.; Pons-Moll, G.; and Theobalt, C. 2019. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proc. CVPR*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*.

Jakab, T.; Gupta, A.; Bilen, H.; and Vedaldi, A. 2018. Unsupervised learning of object landmarks through conditional image generation. In *Proc. NeurIPS*.

Jakab, T.; Gupta, A.; Bilen, H.; and Vedaldi, A. 2019. Learning landmarks from unaligned data using image translation. *arXiv preprint arXiv:1907.02055*.

Johnson, S., and Everingham, M. 2010. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. BMVC*.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proc. CVPR*.

Kocabas, M.; Karagoz, S.; and Akbas, E. 2019. Self-supervised learning of 3d human pose using multi-view geometry. In *Proc. CVPR*.

Kundu, J. N.; Gor, M.; and Babu, R. V. 2019. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proc. AAAI*.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proc. ICCV*.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017a. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*.

Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.-P.; Xu, W.; Casas, D.; and Theobalt, C. 2017b. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*.

Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.-P.; Xu, W.; Casas, D.; and Theobalt, C. 2017c. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*.

Moreno-Noguer, F. 2017. 3d human pose estimation from a single image via distance matrix regression. In *Proc. CVPR*.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *Proc. ECCV*.

Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2012. Reconstructing 3d human pose from 2d image landmarks. In *Proc. ECCV*.

Rhodin, H.; Salzmann, M.; and Fua, P. 2018. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proc. ECCV*.

Rocco, I.; Arandjelovic, R.; and Sivic, J. 2017. Convolutional neural network architecture for geometric matching. In *Proc. CVPR*.

Rogez, G.; Weinzaepfel, P.; and Schmid, C. 2017. Lcr-net: Localization-classification-regression for human pose. In *Proc. CVPR*.

Tung, H.-Y. F.; Harley, A. W.; Seto, W.; and Fragkiadaki, K. 2017. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *Proc. ICCV*.

Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M. J.; Laptev, I.; and Schmid, C. 2017. Learning from synthetic humans. In *Proc. CVPR*.

Wandt, B., and Rosenhahn, B. 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proc. CVPR*.

Wu, J.; Xue, T.; Lim, J. J.; Tian, Y.; Tenenbaum, J. B.; Torralba, A.; and Freeman, W. T. 2016. Single image 3d interpreter network. In *Proc. ECCV*.

Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; and Wang, X. 2018. 3d human pose estimation in the wild by adversarial learning. In *Proc. CVPR*.

Zhang, W.; Liu, Z.; Zhou, L.; Leung, H.; and Chan, A. B. 2017. Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3d human pose estimation. *Image and Vision Computing*.

Zhang, Y.; Guo, Y.; Jin, Y.; Luo, Y.; He, Z.; and Lee, H. 2018. Unsupervised discovery of object landmarks as structural representations. In *Proc. CVPR*.

Zhou, X.; Zhu, M.; Leonardos, S.; and Daniilidis, K. 2016. Sparse representation for 3d shape estimation: A convex relaxation approach. *IEEE transactions on pattern analysis and machine intelligence*.

Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; and Wei, Y. 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proc. ICCV*.