# Self-Supervised 3D Human Pose Estimation via Part Guided Novel Image Synthesis

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

## Abstract

*Camera captured human pose is an outcome of several sources of variation. Performance of supervised 3D pose estimation approaches comes at the cost of dispensing with variations, such as shape and appearance, that may be useful for solving other related tasks. As a result, the learned model not only inculcates task-bias but also dataset-bias because of its strong reliance on the annotated samples, which also holds true for weakly-supervised models. Acknowledging this, we propose a self-supervised learning framework[1] to disentangle such variations from unlabeled video frames. We leverage the prior knowledge on human skeleton and poses in the form of a single part based 2D puppet model, human pose articulation constraints, and a set of unpaired 3D poses. Our differentiable formalization, bridging the representation gap between the 3D pose and spatial part maps, not only facilitates discovery of interpretable pose disentanglement, but also allows us to operate on videos with diverse camera movements. Qualitative results on unseen in-the-wild datasets establish our superior generalization across multiple tasks beyond the primary tasks of 3D pose estimation and part segmentation. Furthermore, we demonstrate state-of-the-art weakly-supervised 3D pose estimation performance on both Human3.6M and MPI-INF-3DHP datasets.*

## 1. Introduction

Analyzing humans takes a central role in computer vision systems. Automatic estimation of 3D pose and 2D part-arrangements of highly deformable humans from monocular RGB images remains an important, challenging and unsolved problem. This ill-posed classical inverse problem has diverse applications in human-robot interaction [53], augmented reality [15], gaming industry, etc.

In a fully-supervised setting [52, 39, 10], the advances in this area are mostly driven by recent deep learning architectures and the collection of large-scale annotated samples. However, unlike 2D landmark annotations, it is very difficult to manually annotate 3D human pose on 2D images. A usual way of obtaining 3D ground-truth (GT) pose annotations is through a well-calibrated in-studio multi-camera setup [20, 50], which is difficult to configure in outdoor environments. This results in a limited diversity in the available 3D pose datasets, which greatly limits the generalization of supervised 3D pose estimation models.

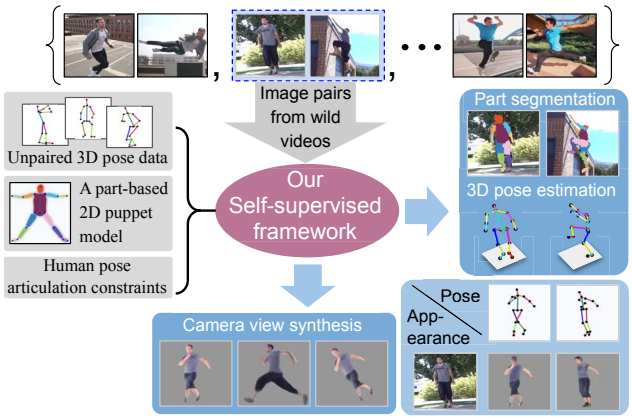


Figure 1. Our self-supervised framework not only produces 3D pose and part segmentation but also enables novel image synthesis via interpretable latent manipulation of the disentangled factors.

To facilitate better generalization, several recent works [7, 46] leverage weakly-supervised learning techniques that reduce the need for 3D GT pose annotations. Most of these works use an auxiliary task such as multi-view 2D pose estimation to train a 3D pose estimator [8, 27]. Instead of using 3D pose GT for supervision, a 3D pose network is supervised with loss functions on multi-view projected 2D poses. To this end, several of these works still require considerable annotations in terms of paired 2D pose GT [7, 57, 42, 28], multi-view images [27] and known camera parameters [46]. Dataset bias still remains a challenge in these techniques as they use paired image and 2D pose GT datasets which have limited diversity. Given the

*Equal contribution.
[1]Project page: http://val.cds.iisc.ac.in/pgp-human/

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00619

ever-changing human fashion and evolving culture, the visual appearance of humans keeps varying and we need to keep updating the 2D pose datasets accordingly.

In this work, we propose a differentiable and modular self-supervised learning framework for monocular 3D human pose estimation along with the discovery of 2D part segments. Specifically, our encoder network takes an image as input and outputs 3 disentangled representations: 1. view-invariant 3D human pose in canonical co-ordinate system, 2. camera parameters and 3. a latent code representing foreground (FG) human appearance. Then, a decoder network takes the above encoded representations, projects them onto 2D and synthesizes FG human image while also producing 2D part segmentation. Here, a major challenge is to disentangle the representations for 3D pose, camera, and appearance. We achieve this disentanglement by training on video frame pairs depicting the same person, but in varied poses. We self-supervise our network with consistency constraints across different network outputs and across image pairs. Compared to recent self-supervised approaches that either rely on videos with static background [45] or work with the assumption that temporally close frames have similar background [22], our framework is robust enough to learn from large-scale in-the-wild videos, even in the presence of camera movements. We also leverage the prior knowledge on human skeleton and poses in the form of a single part-based 2D puppet model, human pose articulation constraints, and a set of unpaired 3D poses. Fig. 1 illustrates the overview of our self-supervised learning framework.

Self-supervised learning from in-the-wild videos is challenging due to diversity in human poses and backgrounds in a given pair of frames which may be further complicated due to missing body parts. We achieve the ability to learn on these wild video frames with a pose-anchored deformation of puppet model that bridges the representation gap between the 3D pose and the 2D part maps in a fully differentiable manner. In addition, the part-conditioned appearance decoding allows us to reconstruct only the FG human appearance resulting in robustness to changing backgrounds.

Another distinguishing factor of our technique is the use of well-established pose prior constraints. In our self-supervised framework, we explicitly model 3D rigid and non-rigid pose transformations by adopting a differentiable parent-relative local limb kinematic model, thereby reducing ambiguities in the learned representations. In addition, for the predicted poses to follow the real-world pose distribution, we make use of an unpaired 3D pose dataset. We interchangeably use predicted 3D pose representations and sampled real 3D poses during training to guide the model towards a plausible 3D pose distribution.

Our network also produces useful part segmentations. With the learned 3D pose and camera representations, we model depth-aware inter-part occlusions resulting in robust part segmentation. To further improve the segmentation beyond what is estimated with pose cues, we use a novel differentiable shape uncertainty map that enables extraction of limb shapes from the FG appearance representation.

We make the following main contributions:

- We present techniques to explicitly constrain the 3D pose by modeling it at its most fundamental form of rigid and non-rigid transformations. This results in interpretable 3D pose predictions, even in the absence of any auxiliary 3D cues such as multi-view or depth.
- We propose a differentiable part-based representation which enables us to selectively attend to foreground human appearance which in-turn makes it possible to learn on in-the-wild videos with changing backgrounds in a self-supervised manner.
- We demonstrate generalizability of our self-supervised framework on *unseen* in-the-wild datasets, such as LSP [23] and YouTube. Moreover, we achieve *state-of-the-art* weakly-supervised 3D pose estimation performance on both Human3.6M [20] and MPI-INF-3DHP [36] datasets against the existing approaches.

## 2. Related Works

Human 3D pose estimation is a well established problem in computer vision, specifically in fully supervised paradigm. Earlier approaches [43, 63, 56, 9] proposed to infer the underlying graphical model for articulated pose estimation. However, the recent CNN based approaches [5, 40, 37] focus on regressing spatial keypoint heat-maps, without explicitly accounting for the underlying limb connectivity information. However, the performance of such models heavily relies on a large set of paired 2D or 3D pose annotations. As a different approach, [26] proposed to regress latent representation of a trained 3D pose autoencoder to indirectly endorse a plausibility bound on the output predictions. Recently, several weakly supervised approaches utilize varied set of auxiliary supervision other than the direct 3D pose supervision (see Table 1). In this paper, we address a more challenging scenario where we consider access to only a set of unaligned 2D pose data to facilitate the learning of a plausible 2D pose prior (see Table 1).

In literature, while several supervised shape and appearance disentangling techniques [3, 34, 33, 13, 49] exist, the available unsupervised pose estimation works (*i.e.* in the absence of multi-view or camera extrinsic supervision), are mostly limited to 2D landmark estimation [11, 22] for rigid or mildly deformable structures, such as facial landmark detection, constrained torso pose recovery etc. The general idea [25, 47, 55, 54] is to utilize the relative transformation between a pair of images depicting a consistent appearance with varied pose. Such image pairs are usually sampled from a video satisfying appearance invariance [22] or synthetically generated deformations [47].

Table 1. Characteristic comparison of our approach against prior weakly-supervised human 3D pose estimation works, in terms of access to direct (paired) or indirect (unpaired) supervision levels.

| Methods | Paired sup. (MV: muti-view) | | | Unpaireed 2D/3D pose Supervision | Sup. for latent to 3D pose mapping |
|---|---|---|---|---|---|
| | MV pair | Cam. extrin. | 2D pose | | |
| Rhodin *et al*. [45] | ✓ | ✓ | ✗ | ✗ | ✓ |
| Kocabas *et al*. [27] | ✓ | ✗ | ✓ | ✗ | ✗ |
| Chen *et al*. [8] | ✓ | ✗ | ✓ | ✗ | ✓ |
| Wandt *et al*. [59] | ✗ | ✗ | ✓ | ✓ | ✗ |
| Chen *et al*. [7] | ✗ | ✗ | ✓ | ✓ | ✗ |
| Ours | ✗ | ✗ | ✗ | ✓ | ✗ |

Beyond landmarks, object parts [32] can infer shape alongside the pose. Part representations are best suited for 3D articulated objects as a result of its occlusion-aware property as opposed to simple landmarks. In general, the available unsupervised part learning techniques [51, 19] are mostly limited to segmentation based discriminative tasks. On the other hand, [61, 41] explicitly leverage the consistency between geometry and the semantic part segments. However, the kinematic articulation constraints are well defined in 3D rather than in 2D [2]. Motivated by this, we aim to leverage the advantages of both non-spatial 3D pose [26] and spatial part-based representation [32] by proposing a novel 2D pose-anchored part deformation model.

## 3. Approach

We develop a differentiable framework for self-supervised disentanglement of 3D pose and foreground appearance from in-the-wild video frames of human activity.

Our self-supervised framework builds on the conventional encoder-decoder architecture (Sec. 3.2). Here, the encoder produces a set of local 3D vectors from an input RGB image. This is then processed through a series of 3D transformations, adhering to the 3D pose articulation constraints to obtain a set of 2D coordinates (camera projected, non-spatial 2D pose). In Sec. 3.1, we define a set of part based representations followed by carefully designed differentiable transformations required to bridge the representation gap between the non-spatial 2D pose and the spatial part maps. This serves three important purposes. First, their spatial nature facilitates compatible input pose conditioning for the fully-convolutional decoder architecture. Second, it enables the decoder to selectively synthesize only FG human appearance ignoring the variations in the background. Third, it facilitates a novel way to encode the 2D joint and part association using a single template puppet model. Finally, Sec. 3.3 describes the proposed self-supervised paradigm which makes use of the pose-aware spatial part maps for simultaneous discovery of 3D pose and part segmentation using image pairs from wild videos.

### 3.1. Joint-anchored spatial part representation

One of the major challenges in unsupervised pose or landmark detection is to map the model-discovered landmarks to the standard landmark conventions. This is essential to facilitate the subsequent task-specific pipelines, which expect the input pose to follow a certain convention. Prior works [45, 30] rely on paired supervision to learn this mapping. In absence of such supervision, we aim to encode this convention in a *canonical part dictionary* where the association of 2D joints with respect to the body parts is extracted from a *single* manually annotated puppet template (Fig. 2C, top panel). This can be interpreted as a 2D human puppet model, which can approximate any human pose deformation via independent spatial transformation of body parts while keeping intact the anchored joint associations.

**Canonical maps.** We extract *canonical part maps*, $\{\phi_c^{(l)}\}_{l=1}^{L}$ (here, $l$: limb index and $L$: total number of limbs or parts), where we perform erosion followed by Gaussian blurring of binary part segments to account for the associated *shape* uncertainty (*i.e*. body shape or apparel shape variations). We represent $\phi_c^{(l)} : \mathcal{U} \rightarrow [0, 1]$, where $\mathcal{U} \in \mathbb{N}^2$ is the space of spatial indices. In addition, we also extract *canonical shape uncertainty maps* $\{\psi_c^{(l)}\}_{l=1}^{L}$ to specifically highlight only the uncertain regions (Fig. 2C, bottom panel). The two anchored joint locations for each limb $l$ and its corresponding part map $\phi_c^{(l)}$ are denoted as $r_c^{l(j_1)}, r_c^{l(j_2)} \in \mathcal{U}$, except for the *torso* which is represented using 4 joints.

**Part deformation model.** For a given 2D pose $q \in \mathbb{R}^{2J}$ with $J$ being the total number of joints, *part-wise pose maps* are obtained as independent spatial-transformations of the *canonical part maps*, *i.e*. $\phi_p^{(l)} = \mathcal{S}^{(l)} \circ \phi_c^{(l)}$. Here, $\mathcal{S}^{(l)}$ represents an affine transformation of the spatial indices $u \in \mathcal{U}$, whose rotation, scale, and translation parameters are obtained as a function of $(q^{l(j_1)}, q^{l(j_2)}, r_c^{l(j_1)}, r_c^{l(j_2)})$, where $q^{l(j_1)}$, $q^{l(j_2)}$ denote the joint locations associated with the limb $l$ in pose $q$. Similarly, we also compute the *part-wise shape uncertainty maps* as $\psi_p^{(l)} = \mathcal{S}^{(l)} \circ \psi_c^{(l)}$. Note that, $\{\phi_p^{(l)}\}_{l=1}^{L}$ and $\{\psi_p^{(l)}\}_{l=1}^{L}$ are unaware of inter-part occlusion in the absence of limb-depth information. Following this, we obtain single-channel maps (see Fig. 2D), *i.e*.

**a)** *shape uncertainty map* as $w_{unc} = \max_l \psi_p^{(l)}$, and

**b)** *single-channel FG-BG map* as $w_{fg} = \max_l \phi_p^{(l)}$.

The above formalization bridges the representation gap between the raw joint locations, $q$ and the output spatial maps $\phi_p$, $w_{fg}$, and $w_{unc}$, thereby facilitating them to be used as differentiable spatial maps for the subsequent self-supervised learning.

**Depth-aware part segmentation.** For 3D deformable objects, a reliable 2D part segmentation can be obtained with the help of following attributes, *i.e*. a) 2D skeletal

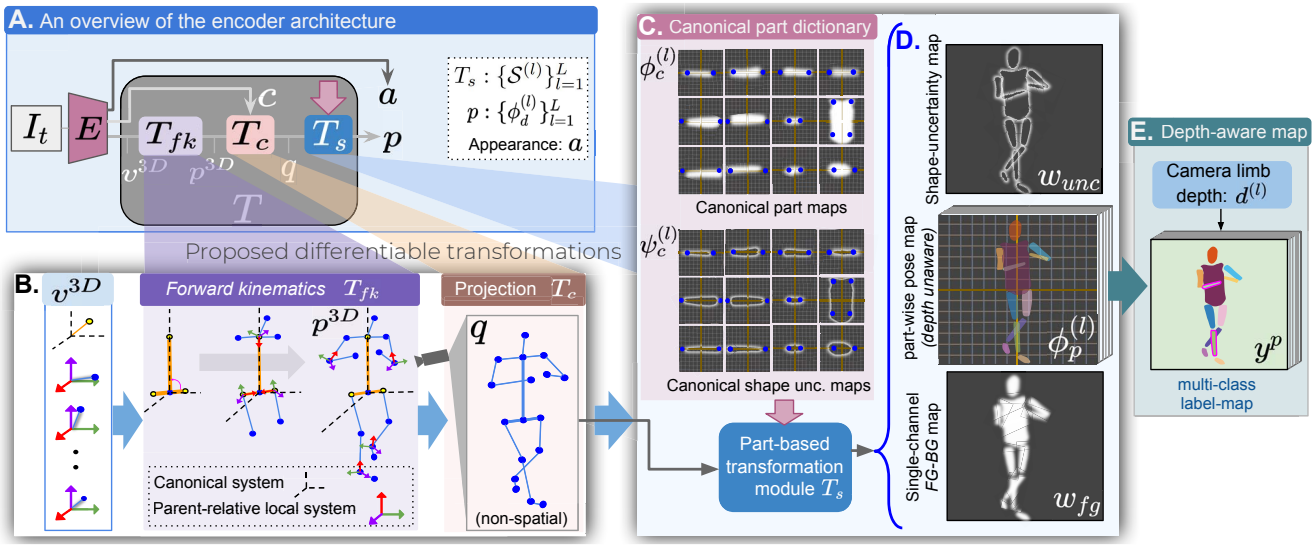


Figure 2. **A.** An overview of the full encoder module. **B.** Transforming the parent-relative local 3D vectors $v_{3D}$ to camera projected 2D pose $q$. **C.** The template 2D puppet model. **D.** Puppet imitates the pose in $q$. **E.** Image independent depth-aware part segmentation.

pose, b) part-shape information, and c) knowledge of inter-part occlusion. Here, the 2D skeletal pose and the knowledge of inter-part occlusion can be extracted by accessing camera transformation of the corresponding 3D pose representation. Let, the depth of the 2D joints in $q$ with respect to the camera be denoted as $q_d^{l(j_1)}$ and $q_d^{l(j_2)}$. We obtain a scalar depth value associated with each limb $l$ as, $d^{(l)} = (q_d^{l(j_1)} + q_d^{l(j_2)})/2$. We use these depth values to alter the strength of depth-unaware *part-wise pose maps*, $\phi_p^{(l)}$ at each spatial location, $u \in \mathcal{U}$ by modulating the strength of part map intensity as being inversely proportional to the depth values. This is realized in the following steps:

**a)** $\phi_d^{(l)}(u) = softmax_{l=1}^{L}(\phi_p^{(l)}(u)/d^{(l)})$,

**b)** $\phi_d^{(L+1)}(u) = 1 - \max_{l=1}^{L} \phi_d^{(l)}(u)$, and

**c)** $\bar{\phi}_d^{(l)}(u) = softmax_{l=1}^{L+1}(\phi_d^{(l)}(u))$.

Here, $(L+1)$ indicates the spatial-channel dedicated for the background. Additionally, a non-differentiable 2D part-segmentation map (see Fig. 2E) is obtained as,

$$y^p(u, l) = \mathbb{1}(l = \text{argmax}_{l=1}^{L+1} \bar{\phi}_d^{(l)}(u)).$$

### 3.2. Self-supervised pose network

The architecture for self-supervised pose and appearance disentanglement consists of a series of pre-defined differentiable transformations facilitating discovery of a constrained latent pose representation. As opposed to imposing learning based constraints [14], we devise a way around where the 3D pose articulation constraints (*i.e.* knowledge of joint connectivity and bone-length) are directly applied via structural means, implying guaranteed constraint imposition.

**a) Encoder network.** As shown in Fig. 2A, the encoder $E$ takes an input image $I$ and outputs three disentangled factors, a) a set of local 3D vectors: $v^{3D} \in \mathbb{R}^{3J}$, b) camera parameters: $c$, and c) a FG appearance: $a \in \mathbb{R}^{H \times W \times \text{Ch}}$.

As compared to spatial 2D geometry [45, 32], discovering the inherent 3D human pose is a highly challenging task considering the extent of associated non-rigid deformation, and rigid camera variations [2, 29]. To this end, we define a canonical coordinate system $C$, where *face-vector* of the skeleton is canonically aligned along the +ve X-axis, thus making it completely view-invariant. Here, the *face-vector* is defined as the perpendicular direction of the plane spanning the neck, left-hip and right-hip joints. As shown in Fig. 2B, in $v^{3D}$, except the pelvis, neck, left-hip and right-hip, all other joints are defined at their respective parent relative local coordinate systems (*i.e.* parent joint as the origin with axis directions obtained by performing Gram-Schmidt orthogonalization of the parent-limb vector and the *face-vector*). Accordingly, we define a recursive forward kinematic transformation $T_{fk}$ to obtain the canonical 3D pose from the local limb vectors, *i.e.* $p^{3D} = T_{fk}(v^{3D})$, which accesses a constant array of limb length magnitudes [68].

Here, the camera extrinsics, $c$ (3 rotation angles and 3 restricted translations ensuring that the camera-view captures all the skeleton joints in $p^{3D}$) is obtained at the encoder output, whereas a fixed perspective camera projection is applied to obtain the final 2D pose representation, *i.e.* $q = T_c(p^{3D})$. A part based deformation operation on this 2D pose $q$ (Sec. 3.1) is shown as $p = T_s(q, d^{(l)})$, where $T_s : \{\mathcal{S}^{(l)}\}_{l=1}^{L}$ and $p : \{\phi_d^{(l)}\}_{l=1}^{L}$ (following the depth aware operations on $\phi_p^{(l)}$). Finally, $T$ denotes the entire series of differentiable transformations, *i.e.* $T_s \circ T_c \circ T_{fk}$, as shown in Fig. 2A. Here, $\circ$ denotes composition operation.

**b) Decoder network.** The decoder takes a concatenated representation of the FG appearance, $a$ and pose, $p$ as input to obtain two output maps, i) a reconstructed image $\hat{I}$,

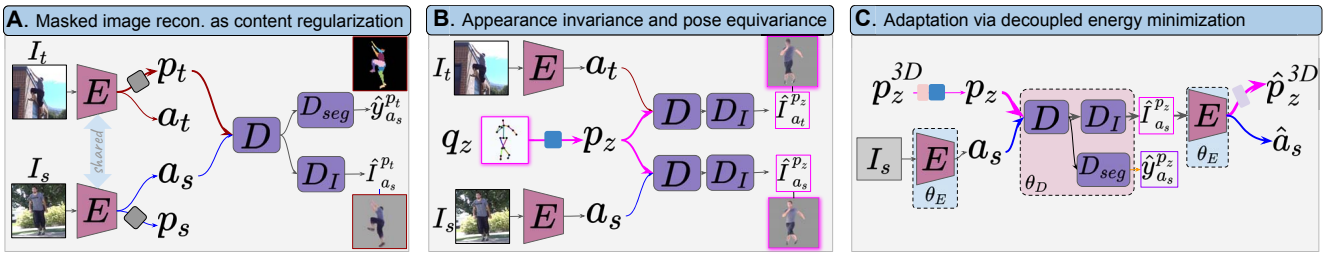


Figure 3. An overview of data-flow pipelines for the proposed self-supervised objectives. Images close to the output heads show the network output for the given set of exemplar input tuple. Here, the color of transformation blocks are consistent with Fig. 2A.

and ii) a predicted part segmentation map $\hat{y}$ via a bifurcated CNN decoder (see Fig. 3A). The common decoder branch, $D$ consists of a series of up-convolutional layers conditioned on the spatial pose map $p$ at each layer's input (*i.e.* multi-scale pose conditioning). Whereas, $D_I$ and $D_{seg}$ follow up-convolutional layers to their respective outputs.

### 3.3. Self-supervised training objectives

The prime design principle of our self-supervised framework is to leverage the interplay between the pose and appearance information by forming paired input images of either consistent pose or appearance.

Given a pair of source and target image, $(I_s, I_t)$, sampled from the same video *i.e.* with consistent FG appearance, the shared encoder extracts their respective pose and appearance as $(p_s, p_t)$ and $(a_s, a_t)$ (see Fig. 3A). We denote the decoder outputs as $(\hat{I}^{p_t}_{a_s}, \hat{y}^{p_t}_{a_s})$ while the decoder takes in pose, $p_t$ with appearance, $a_s$ (this notation is consistent in later sections). Here, $\hat{I}^{p_t}_{a_s}$ is expected to depict the person in the target pose $p_t$. Such a cross-pose transfer setup is essential to restrict leakage of pose information through the appearance. Whereas, the low dimensional bottleneck of $v^{3D}$ followed by the series of differentiable transformations prevents leakage of appearance through pose [22].

To effectively operate on wild video frames (*i.e.* beyond the in-studio fixed camera setup [45, 22]), we aim to utilize the pose-aware, spatial part representations as a means to disentangle the FG from BG. Thus, we plan to reconstruct $\hat{I}^{p_t}_{a_s}$ with a constant BG color $BG_c$ and segmented FG appearance (see Fig. 3A). Our idea stems from the concept of co-saliency detection [64, 18], where the prime goal is to discover the common or salient FG from a given set of two or more images. Here, the part appearances belonging to the model predicted part regions have to be consistent across $I_s$ and $I_t$ for a successful self-supervised pose discovery.

**Access to unpaired 3D/2D pose samples.** We denote $p^{3D}_z$ and $q_z = T_c(p^{3D}_z)$ as a 3D pose and its projection (via random camera), sampled from an *unpaired* 3D pose dataset $\mathcal{D}_z$, respectively. Such samples can be easily collected without worrying about BG or FG diversity in the corresponding camera feed (*i.e.* a single person activity). We use these samples to further constrain the latent space towards realizing a plausible 3D pose distribution.

**a) Image reconstruction objective.** Unlike [45, 22], we do not have access to the corresponding ground-truth representation for the predicted $\hat{I}^{p_t}_{a_s}$, giving rise to an increased possibility of producing degenerate solutions or *mode-collapse*. In such cases, the model focuses on fixed background regions as the common region between the two images, specifically for in-studio datasets with limited BG variations. One way to avoid such scenarios is to select image pairs with completely diverse BG (*i.e.* select image pairs with high L2 distance in a sample video-clip).

To explicitly restrict the model from inculcating such BG bias, we incorporate content-based regularization. An uncertain pseudo FG mask is used to establish a one-to-one correspondence between the reconstructed image $\hat{I}^{p_t}_{a_s}$ and the target image $I_t$. This is realized through a spatial-mask, $m_{sal}$ which highlights the salient regions (applicable for any image frame) or regions with diverse motion cues (applicable for frames captured in a fixed camera). We formulate a pseudo (*uncertain*) reconstruction objective as,

$$\mathcal{L}^u_I = (1 - \hat{y}^{p_t}_{a_s}(L+1) + \beta m^{I_t}_{sal}) \otimes |\hat{I}^{p_t}_{a_s} - I_t|$$

Here, $\otimes$ denotes pixel-wise weighing and $\beta$ is a balancing hyperparameter. Note that, the final loss is computed as average over all the spatial locations, $u \in \mathcal{U}$. This loss enforces a self-supervised consistency between the pose-aware part maps, $\hat{y}^{p_t}_{a_s}$ and the salient common FG to facilitate a reliable pose estimate.

As a novel direction, we utilize $q_z \sim \mathcal{D}_z$ to form a pair of image predictions $(\hat{I}^{p_z}_{a_s}, \hat{I}^{p_z}_{a_t})$ following simultaneous appearance invariance and pose equivariance (see Fig. 3B). Here, a *certain* reconstruction objective is defined as,

$$\mathcal{L}^c_I = w^{p_z}_{fg} \otimes |\hat{I}^{p_z}_{a_s} - \hat{I}^{p_z}_{a_t}| + (1 - w^{p_z}_{fg}) \otimes |\hat{I}^{p_z}_{a_s} - BG_c|$$

**b) Part-segmentation objective.** Aiming to form a consistency between the true pose $p_z$ and the corresponding part segmentation output, we formulate,

$$\mathcal{L}_{seg} = (1 - w_{unc}) \otimes CE(\hat{y}^{p_z}_{a_s}, y^{p_z}) + w_{unc} \otimes SE(\hat{y}^{p_z}_{a_s})$$

Here, *CE*, and *SE* denote the pixel-wise cross-entropy and self-entropy respectively. Moreover, we confidently enforce segmentation loss with respect to one-hot map $y^{p_z}$ (Sec. 3.1) only at the *certain* regions, while minimizing the

Shanon's entropy for the regions associated with *shape uncertainty* as captured in $w_{unc}$. Here, the limb depth required to compute $y^{p_z}$ is obtained from $\hat{p}_z^{3D} = E_p(\hat{I}_{a_s}^{p_z})$ (Fig. 3C).

In summary, the above self-supervised objectives form a consistency among $p$, $\hat{y}$, and $\hat{I}$;

**a)** $\mathcal{L}_I^u$ enforces consistency between $\hat{y}$ and $\hat{I}$,

**b)** $\mathcal{L}_I^c$ enforces consistency between $p$ (via $w_{fg}$) and $\hat{I}$,

**c)** $\mathcal{L}_{seg}$ enforces consistency between $p$ (via $y^{p_z}$) and $\hat{y}$.

However, the model inculcates a discrepancy between the predicted pose and the true pose distributions. It is essential to bridge this discrepancy as $\mathcal{L}_I^c$ and $\mathcal{L}_{seg}$ rely on true pose $q_z = T_c(p_z^{3D})$, whereas $\mathcal{L}_I^u$ relies on the predicted pose $q_t$. Thus, we employ an adaptation strategy to guide the model towards realizing a plausible pose prediction.

**c) Adaptation via energy minimization.** Instead of employing an ad-hoc adversarial discriminator [7, 62], we devise a simpler yet effective decoupled energy minimization strategy [16, 21]. We avoid a direct encoder-decoder interaction during gradient back-propagation, by updating the encoder parameters, while freezing the decoder parameters and vice-versa. However, this is performed while enforcing a reconstruction loss at the output of the *secondary* encoder in a cyclic auto-encoding scenario (see Fig. 3C). The two energy functions are formulated as $\mathcal{L}_{p_z^{3D}} = |p_z^{3D} - \hat{p}_z^{3D}|$ and $\mathcal{L}_{a_s} = |a_s - \hat{a}_s|$, where $\hat{p}_z^{3D} = E_p(\hat{I}_{a_s}^{p_z})$ and $\hat{a}_s = E_a(\hat{I}_{a_s}^{p_z})$.

The decoder parameters are updated to realize a faithful $\hat{I}_{a_s}^{p_z}$, as the frozen encoder expects $\hat{I}_{a_s}^{p_z}$ to match its input distribution of real images (*i.e.* $I_s$) for an effective energy minimization. Here, the encoder can be perceived as a frozen energy network as used in energy-based GAN [66]. A similar analogy applies while updating the encoder parameters with gradients from the frozen decoder. Each alternate energy minimization step is preceded by an overall optimization of the above consistency objectives, where both encoder and decoder parameters are updated simultaneously (see Algo. 1).

$\theta_E$: Trainable parameters of the Encoder $E$
$\theta_D$: Trainable parameters of the Decoder
(includes $D$, $D_I$, and $D_{seg}$)
**for** *iter* $<$ *MaxIter* **do**
  **if** *iter* $(\mathrm{mod}\ 2) \neq 0$ **then**
    Update $\theta_E$ by optimizing $\mathcal{L}_{p_z^{3D}}$ and $\mathcal{L}_{a_s}$ in separate *Adagrad* optimizers on frozen $\theta_D$.
  **else**
    Update $\theta_D$ by optimizing $\mathcal{L}_{p_z^{3D}}$ and $\mathcal{L}_{a_s}$ in separate *Adagrad* optimizers on frozen $\theta_E$.
  **end**
  Update $(\theta_E, \theta_D)$ by optimizing $\mathcal{L}_I^u$, $\mathcal{L}_I^c$, and $\mathcal{L}_{seg}$ in separate *Adagrad* optimizers.
**end**

**Algorithm 1:** Self-supervised learning with the proposed adaptation via decoupled energy minimization.

# 4. Experiments

We perform a thorough experimental analysis to establish the effectiveness of our proposed framework on 3D pose estimation, part segmentation and novel image synthesis tasks, across several datasets beyond the in-studio setup.

**Implementation details.** We employ an ImageNet trained *Resnet-50* architecture [17] as the base CNN for the encoder $E$. We first bifurcate it into two CNN branches dedicated to pose and appearance, then the pose branch is further bifurcated into two multi-layer fully-connected networks to obtain the local pose vectors $v^{3D}$ and the camera parameters $c$. While training, we use separate AdaGrad optimizers [12] for each loss term at alternate training iterations. We perform appearance (color-jittering) and pose augmentations (mirror flip and inplane rotation) selectively for $I_t$ and $I_s$ conceding their invariance effect on $p_t$ and $a_s$ respectively.

**Datasets.** We train the *base-model* on image pairs sampled from a mixed set of video datasets *i.e.* Human3.6M [20] (H3.6M) and an in-house collection of in-the-wild YouTube videos (YTube). As opposed to the in-studio H3.6M images, the YTube dataset constitutes a substantial diversity in apparels, action categories (dance forms, parkour stunts, etc.), background variations, and camera movements. The raw video frames are pruned to form the suitable image pairs after passing them through an off-the-shelf person-detector [44]. We utilize an unsupervised saliency detection method [70] to obtain $m_{sal}$ for the wild YTube frames, whereas for samples from H3.6M $m_{sal}$ is obtained directly through the BG estimate [45]. Further, LSP [31] and MPI-INF-3DHP [36] (3DHP) datasets are used to evaluate generalizability of our framework. We collect the unpaired 3D pose samples, $q_z$ from MADS [65] and CMU-MoCap [1] dataset keeping a clear domain gap with respect to the standard datasets chosen for benchmarking our performance.

## 4.1. Evaluation on Human3.6M

Inline with the prior arts [7, 45], we evaluate our 3D pose estimation performance in the standard protocol-II setting (*i.e.* with scaling and rigid alignment). We experimented on 4 different variants of the proposed framework with increasing degrees of supervision levels. The base model in absence of any paired supervision is regarded as *Ours(unsup)*. In presence of multi-view information (with camera extrinsics), we finetune the model by enforcing consistent canonical pose $p^{3D}$ and camera shift for multi-view image pairs, termed as *Ours(multi-view-sup)*. Similarly, finetuning in presence of a direct supervision on the corresponding 2D pose GT is regarded as *Ours(weakly-sup)*. Lastly, finetuning in presence of a direct 3D pose supervision on 10% of the full training set is referred to as *Ours(semi-sup)*. Table 2 depicts our superior performance against the prior arts in their respective supervision levels.

Table 2. Comparison of 3D pose estimation results on Human3.6M. Comparable metrics of fully-supervised methods are included for reference. Our approach achieves state-of-the art performance while brought to the same supervision-level (divided by horizontal lines) of *Full-2D* (row no. 4-8) or *Mutli-view* (row no. 9-10). Moreover, *Ours(semi-sup)* achieves comparable performance against the prior fully supervised approaches.

| No. | Protocol-II | Supervision | Avg. MPJPE(↓) |
|---|---|---|---|
| 1. | Zhou *et al.* [69] | Full-3D | 106.7 |
| 2. | Chen *et al.* [6] | Full-3D | 82.7 |
| 3. | Martinez *et al.* [35] | Full-3D | 52.1 |
| 4. | Wu *et al.* [60] | Full-2D | 98.4 |
| 5. | Tung *et al.* [57] | Full-2D | 97.2 |
| 6. | Chen *et al.* [7] | Full-2D | 68.0 |
| 7. | Wandt *et al.* [59] | Full-2D | 65.1 |
| 8. | *Ours(weakly-sup)* | Full-2D | **62.4** |
| 9. | Rhodin *et al.* [45] | Multi-view | 98.2 |
| 10. | *Ours(multi-view-sup)* | Multi-view | **85.8** |
| 11. | *Ours(unsup)* | **No sup.** | 99.2 |
| 12. | *Ours(semi-sup)* | 10%-3D | **50.8** |

Table 3. Ablation analysis, highlighting importance of various constraints and regularization in the proposed self-supervised 3D pose estimation framework. (Qualitative results in Fig. 5B)

| Method (unsup.) | MPJPE(↓) on Human3.6M | 3DPCK(↑) on MPI-3DHP |
|---|---|---|
| *Ours(unsup)* w/o $T_{fk} \circ T_c$ | 126.8 | 51.7 |
| *Ours(unsup)* w/o $q_z \sim \mathcal{D}_z$ | 178.9 | 40.3 |
| *Ours(unsup)* w/o $m_{sal}$ | 189.4 | 35.7 |
| *Ours(unsup)* | 99.2 | 77.4 |

## 4.2. Evaluation on MPI-INF-3DHP

With our framework, we also demonstrate a higher level of cross-dataset generalization, thereby minimizing the need for finetuning on novel unseen datasets. The carefully devised constraints at the intermediate pose representation are expected to restrict the model from producing implausible poses even when tested in unseen environments. To evaluate this, we directly pass samples of MPI-INF-3DHP [36] (3DHP) test-set through *Ours(weakly-sup)* model trained on YTube+H3.6M and refer it as unsupervised transfer, denoted as -3DHP in Table 4. Further, we finetune the *Ours(weakly-sup)* on 3DHP dataset at three supervision levels, a) no supervision, b) full 2D pose supervision, and c) 10% 3D pose supervision as reported in Table 4, at rows 10, 7, and 11 respectively. The reported metrics clearly highlight our superiority against the prior arts.

## 4.3. Ablation study

We evaluate the effectiveness of the proposed local vector representation followed by the forward kinematic transformations, against a direct estimation of the 3D joints in camera coordinate system in the presence of appropriate bone length constraints [7]. As reported in Table 3, our disentanglement of camera from the view-invariant canonical pose shows a clear superiority as a result of using the 3D pose articulation constraints in the most fundamental form. Besides this, we also perform ablations by removing $q_z$ or $m_{sal}$ from the unsupervised training pipeline. As shown in Fig. 5B, without $q_z$ the model predicts implausible part arrangements even while maintaining a roughly consistent FG silhouette segmentation. However, without $m_{sal}$, the model renders a plausible pose on the BG area common between the image pairs, as a degenerate solution.

Table 4. 3D pose estimation on 3DHP. Here, 2nd column denotes whether the approach uses 3DHP samples (paired or unpaired) while training. And the 3rd column specifies the supervision level.

| No. | Method | Trainset | 3DHP Sup. | PCK (↑) | AUC (↑) | MPJPE (↓) |
|---|---|---|---|---|---|---|
| 1. | Mehta *et al.* [38] | +3DHP | Full-3D | 76.6 | 40.4 | 124.7 |
| 2. | Rogez *et al.* [48] | +3DHP | Full-3D | 59.6 | 27.6 | 158.4 |
| 3. | Zhou *et al.* [67] | +3DHP | Full-2D | 69.2 | 32.5 | 137.1 |
| 4. | HMR [24] | +3DHP | Full-2D | 77.1 | 40.7 | 113.2 |
| 5. | Yang *et al.* [62] | +3DHP | Full-2D | 69.0 | 32.0 | - |
| 6. | Chen *et al.* [7] | +3DHP | Full-2D | 71.7 | 36.3 | - |
| 7. | *Ours(weakly-sup)* | +3DHP | Full-2D | **84.6** | **60.8** | **93.9** |
| 8. | Chen *et al.* [7] | -3DHP | - | 64.3 | 31.6 | - |
| 9. | *Ours(weakly-sup)* | -3DHP | - | 82.1 | 56.3 | 103.8 |
| 10. | *Ours(weakly-sup)* | +3DHP | **No sup.** | **83.2** | **58.7** | **97.6** |
| 11. | *Ours(semi-sup)* | +3DHP | 10%-3D | 86.3 | 62.0 | 74.1 |

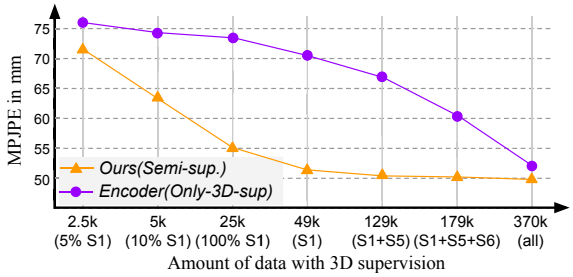


Figure 4. 3D pose estimation on H3.6M as a function of the amount of training supervision. *Ours(semi-sup)* shows faster transferability as compared to the fully supervised counterpart.

As an ablation of the semi-supervised setting, (see Fig. 4), we train the proposed framework on progressively increasing amount of 3D pose supervision alongside the unsupervised learning objective. Further, we perform the same for the Encoder network without the unsupervised objectives (thus, discarding the decoder networks) and term it as *Encoder(Only-3D-sup)*. The plots in Fig. 4 clearly highlight our reduced dependency on the supervised data implying graceful and faster transferability.

## 4.4. Evaluation of part segmentation

For evaluation of part-segmentation, we standardize the ground-truth part conventions across both LSP [23] and H3.6M datasets via SMPL model fitting [58, 31]. This convention roughly aligns with the part-puppet model used in Fig. 2C, thereby maintaining a consistent part to joint association. Note that, $w_{unc}$ is supposed to account for the ambiguity between the puppet-based shape-unaware segmen-

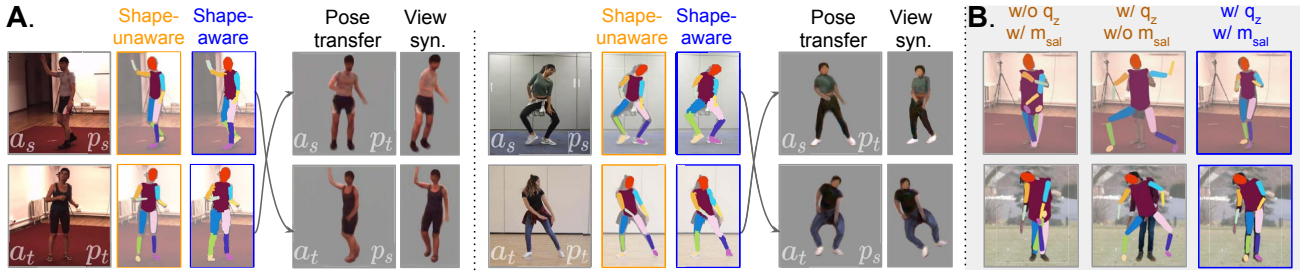


Figure 5. **A.** Novel image synthesis via latent manipulation of $a$, $p$ and $c$. It also shows the effect of independent non-rigid (*pose-transfer*) and rigid (*view-syn.*) variations as a result of explicit disentanglement. Notice the corresponding shape-unaware (puppet deformation) and shape-aware part-segmentation results. **B.** Qualitative analysis of ablations showing importance of $q_z$ and $m_{sal}$ (refer Sec. 4.3).

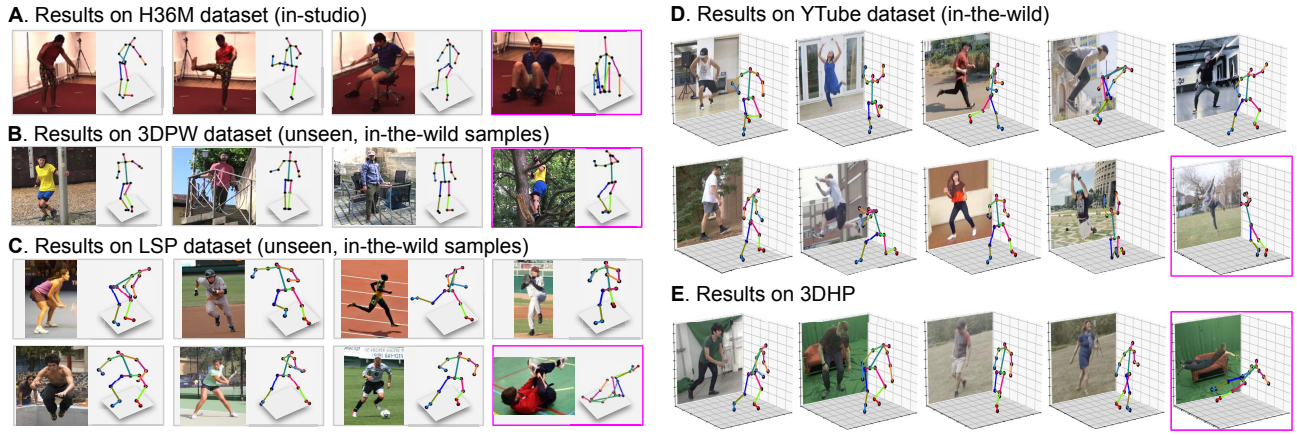


Figure 6. Qualitative results on 5 different datasets. Failure cases are highlighted in magenta which specifically occur in presence of multi-level inter-limb occlusion (see LSP failure case) and very rare, athletic poses (see YTube failure case). However, the model faithfully attends to single-level occlusions, enabled by the depth-aware part representation.

Table 5. Segmentation results on F1 metric (↑) for LSP dataset.

| Method | Pose Sup. | FG vs BG | FG Parts |
|---|---|---|---|
| SMPLify [4] | Full-2D + *SMPL-fitting* | 0.88 | 0.64 |
| HMR [24] | Full-2D + *SMPL-fitting* | 0.86 | 0.59 |
| *Ours(weakly-sup)* | Full-2D (no *SMPL*) | 0.84 | 0.56 |
| *Ours(unsup)* | No sup. (no *SMPL*) | 0.78 | 0.47 |

tation against the image dependent shape-aware segmentation. In Fig. 5A, we show the effectiveness of this design choice, where the shape-unaware segmentation is obtained at $y^p$ after depth-based part ordering, and the corresponding shape-aware segmentation is obtained at $\hat{y}$ output. Further, quantitative comparison of part segmentation is reported in Table 5. We achieve comparable results against the prior arts, in absence of additional supervision.

### 4.5. Qualitative results

To evaluate the effectiveness of the disentangled factors beyond the intended primary task of 3D pose estimation and part segmentation, we manipulate them to analyze their effect on the decoder synthesized output image. In *pose-transfer*, pose obtained from an image is transferred to the appearance of another. However, in *view-syn.*, we randomly vary the camera extrinsic values in $c$. The results shown in Fig. 5A are obtained from *Ours(unsup)* model, which is trained on the mixed YTube+H3.6 dataset. This demonstrates the clear disentanglement of pose and appearance. Fig. 6 depicts qualitative results for the primary 3D pose estimation and part segmentation tasks using *Ours(weakly-sup)* model as introduced in Sec. 4.1. In Fig. 6B, we show results on the unseen LSP dataset, where the model has not seen this dataset even during self-supervised training. A consistent performance on such unseen dataset further establishes generalizability of the proposed framework.

## 5. Conclusion

We proposed a self-supervised 3D pose estimation method that disentangles the inherent factors of variations via part guided human image synthesis. Our framework has two prominent traits. First, effective imposition of both human 3D pose articulation and joint-part association constraint via structural means. Second, usage of depth-aware part based representation to specifically attend to the FG human resulting in robustness to changing backgrounds. However, extending such a framework for multi-person or partially visible human scenarios remains an open challenge.

**Acknowledgements.** This project is supported by a Indo-UK Joint Project (DST/INT/UK/P-179/2017), DST, Govt. of India and a Wipro PhD Fellowship (Jogendra).

## References


[1] CMU graphics lab motion capture database. available: http://mocap.cs.cmu.edu/. 6

[2] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR*, 2015. 3, 4

[3] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. 2

[4] 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. 8

[5] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016. 2

[6] Ching-Hang Chen and Deva Ramanan. 3D human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2017. 7

[7] 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, 3, 6, 7

[8] 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, 3

[9] Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NeurIPS*, 2014. 2

[10] 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

[11] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *NeurIPS*, 2017. 2

[12] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011. 6

[13] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 2

[14] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3D representations. In *CVPR*, 2019. 4

[15] 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*, 2010. 1

[16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 6

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

[18] Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Unsupervised cnn-based co-saliency detection with graphical optimization. In *ECCV*, 2018. 5

[19] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. SCOPS: Self-supervised co-part segmentation. In *CVPR*, 2019. 3

[20] 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*, 2013. 1, 2, 6

[21] Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In *ICML*, 2017. 6

[22] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NeurIPS*, 2018. 2, 5

[23] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2, 7

[24] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 7, 8

[25] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016. 2

[26] Isinsu Katircioglu, Bugra Tekin, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Learning latent representations of 3D human pose with deep neural networks. *International Journal of Computer Vision*, 2018. 2, 3

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

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

[29] Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and R Venkatesh Babu. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *WACV*, 2019. 4

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

[31] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2d human representations. In *CVPR*, 2017. 6, 7

[32] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *CVPR*, 2019. 3, 4

[33] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017. 2

[34] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 2

[35] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 7

[36] 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. 2, 6, 7

[37] 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*, 2017. 2

[38] 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*, 2017. 7

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

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

[41] David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *CVPR*, 2017. 3

[42] 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

[43] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 2014. 2

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 6

[45] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *ECCV*, 2018. 2, 3, 4, 5, 6, 7

[46] 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

[47] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017. 2

[48] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 7

[49] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018. 2

[50] 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*, 2010. 1

[51] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 3

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

[53] Mikael Svenstrup, Søren Tranberg, Hans Jørgen Andersen, and Thomas Bak. Pose estimation and adaptive robot behaviour for human-robot interaction. In *ICRA*, 2009. 1

[54] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NeurIPS*, 2017. 2

[55] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. 2

[56] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NeurIPS*, 2014. 2

[57] 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, 7

[58] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 7

[59] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In *CVPR*, 2019. 3, 7

[60] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3D interpreter network. In *ECCV*, 2016. 7

[61] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016. 3

[62] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3D human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 6, 7

[63] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 2

[64] Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE transactions on pattern analysis and machine intelligence*, 2016. 5

[65] Weichen Zhang, Zhiguang Liu, Liuyang Zhou, Howard Leung, and Antoni B Chan. Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3D human pose estimation. *Image and Vision Computing*, 2017. 6

[66] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *ICLR*, 2017. 6

[67] 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. 7

[68] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *ECCV*, 2016. 4

[69] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, and Kostas Daniilidis. Sparse representation for 3D shape estimation: A convex relaxation approach. *IEEE transactions on pattern analysis and machine intelligence*, 2016. 7

[70] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014. 6