Unsupervised Cross-Dataset Adaptation via Probabilistic Amodal 3D Human Pose Completion

Jogendra Nath Kundu* Rahul M V* Jay Patravali* R. Venkatesh Babu
Video Analytics Lab, CDS, Indian Institute of Science, Bangalore, India
jogendrak@iisc.ac.in, rmvenkat@andrew.cmu.edu, jaypatravali@gmail.com, venky@iisc.ac.in

Abstract

Despite remarkable success of supervised deep learning models for 3D human pose estimation, performance of such models is mostly limited to constrained laboratory settings. Such models not only exhibit an alarming level of dataset bias, but also fail to operate on unconstrained videos in the presence of external variations such as camera motion, partial body visibility, occlusion, etc. Acknowledging these shortcomings, firstly, we aim to formalize a motion representation learning framework by effectively utilizing both constrained and artificially generated unconstrained video samples for datasets with 3D pose annotation. Without ignoring the inherent uncertainty in pose estimation for the truncated video frames, we devise a novel probabilistic amodal pose completion framework to enable generation of multiple plausible pose-filling outcomes. Secondly, to address dataset bias, the probabilistic amodal framework is re-utilized to design novel self-supervised objectives. This not only enables adaptation of the model to target unannotated datasets (wild YouTube videos) but also encourages learning of generic motion representations beyond the available supervised data even in unconstrained scenarios. Such a training regime helps us achieve state-of-the art performance on unsupervised cross-dataset pose estimation, with a significant improvement in partially-visible unconstrained scenarios.

1. Introduction

Understanding visual content from unconstrained internet videos [23] is an important yet challenging problem. Such videos involve extensive variations in camera motion (such as zooming, panning, translation etc.), which make them highly diverse and challenging for machine-analysis as compared to the standard video datasets for pose-based action recognition or retrieval [44, 10]. Surely, one cannot rely on approaches utilizing low-level motion cues [45] in such wild scenarios due to the inherent difficulty in disentangling the camera motion from the motion elicited by the object of interest. In this work, we focus on extraction of human motion representation from such unconstrained videos. We also aim to improve generalizability of such motion representations across diverse motion categories and in-the-wild datasets beyond the standard laboratory setting [10, 55].

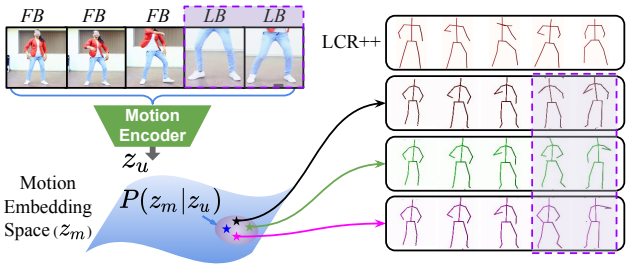

Figure 1. In contrast to deterministic methods such as LCR++ [43], our probabilistic model generates multiple plausible pose sequences (specifically for the invisible body-joints, i.e. the upper-body) for a given unconstrained video with partial-visibility.

A simple approach would be to consider such videos as a set of images [53], which can be directly processed by image-based pose estimation approaches [29, 43]. However, there are many limitations to this approach. **a)** Firstly, datasets with 3D pose [10, 28] information are either limited to laboratory setting or limited in size and diversity. Hence, neural networks trained on such datasets [42, 38, 31, 6] yield impressive performance within the same dataset but do not generalize to unknown motion and camera positions. **b)** Secondly, for images with occlusion or truncated boundary [43] (in absence of full-body (FB) visibility), most of the available pose estimation methods fail as a result of their dependency on intermediate 2D joint estimation. Though certain recent approaches [43] address this by hallucinating a plausible 3D pose considering visibility of the upper-body (UB), they do not take into account uncertainty in the lower-body pose and other diverse scenarios such as only head and only lower-body (LB) visibility. As a result, such approach

*Equal contribution

lead to temporally inconsistent results with drastic transitions such as *FB-UB-LB* or *FB-LB-FB* in an unconstrained scenario. **c)** Thirdly, treating a video as a set of images restrain the model from exploiting temporal coherency, which is an important motion cue.

Aiming to effectively address the above limitations we employ the following design choices. Motivated from recent works focusing on learning a generative pose embedding representation [18, 19], we plan to model a latent skeleton representation called *pose-vec*, after disentangling the camera-view and translation to learn a view agnostic representation. A CNN is trained to directly regress the *pose-vec* latent code from a given input image independent of the camera view. Devoid of the dependency on intermediate 2D pose estimation [42, 29], this approach is easily extendable to truncated images. However, uniqueness of 3D pose for such samples do not carry over as a result of uncertainty in 3D joint locations belonging to the invisible body-parts. We aim to explicitly model this uncertainty by formalizing it as a conditional distribution over three important factors viz. a) position of visible joints at that time instance, b) full-body motion dynamics. By casting it as a probabilistic framework, we plan to predict multiple plausible pose-filling outcomes for a given video sequence with minimal full-body frames in an unconstrained setting (see Fig. 1). This is realized for datasets with 3D pose annotations, by experimentally simulating random videographic variation from videos collected using multi-view static-cameras [10, 28]. To accomplish this, a structured simulation pipeline is designed by analyzing the factors involved in a video capturing process [23]. However, it is not straightforward for datasets beyond the lab environment in the absence of 3D pose annotation.

Researchers have explored several self-supervised [4, 37, 33] and unsupervised [52, 4] approaches to improve generalization of available models to unseen environments [21, 20]. In this paper, we plan to formalize a self-supervised learning framework for human pose estimation particularly targetting unconstrained videos with no access to pose annotations (e.g videos collected from Youtube). By effectively exploiting the temporal coherency, we also seek to obtain a unified model which can yield highly generalizable motion representation. For instance, motion representation of a dance performer from a TV broadcasted video (with frequent scene-change and other camera movements) should semantically align with the representation from any other video consisting some other dancer imitating the exact same dance steps. The proposed pose estimation pipeline is specially designed to address this as a result of the *pose-vec* being absolutely invariant to diverse videography shifts (even shifts of camera-views).

As a novel direction, we designed several self-supervision strategies by exploiting the videography augmentations in a very judicious manner. Motivated by self-supervised approaches which configure correspondences between image pairs by simulating know synthetic warps [34, 48, 12, 41], we plan to use the videography augmentations to simulate similar correspondences among full-body frames between two video-clips, which would have the same pose representation. Though such a strategy seems effective for full-body frames from a target unannotated dataset, it fails to form such correspondence sets for frames with partial-body visibility. As discussed before, instead of ignoring the uncertainty in pose for such truncated frames, we formalize a workaround by introducing a random-vector, which can effectively model this uncertainty in a probabilistic framework [18, 57]. Following this, we acquire an absolutely certain correspondence pair even for such truncated frames without disturbing the generative probabilistic setup. The above strategies greatly enrich our self-supervised approach by enabling the model to learn pose or motion representations that are robust to diverse unconstrained scenarios.

To clearly highlight merits of the proposed learning pipeline, we evaluated our model against state-of-the-art approaches on the pose estimation task. Through extensive experiments on multiple annotated and unannotated datasets, we demonstrate state-of-the-art transferable representation against the previous arts.

2. Related work

Representation learning. Classical auto-encoder framework [50, 9, 2] has been widely used as a base setup in variety of tasks to extract meaningful semantic information from raw unlabelled input data [17]. Recently, self-supervised learning has emerged as a new direction in unsupervised representation learning [22, 37, 54]. Furthermore, representation learning from videos has also gained substantial attention considering the fact that human learns crucial visual representation by seeing moving objects in the environment [36, 30, 52]. Other set of works focus on cross modal information for representation learning [35, 11].

Amodal completion. Humans have the ability to perceive invisible or occluded object parts as a result of general object representations in the brain [32]. Amodal completion of object instances [24] from static images has been studied extensively in recent past [7, 58, 24, 13]. However to the best of our knowledge ours is the first work to extend the concept of amodal perception for temporally coherent motion dynamics.

Human pose estimation from monocular image. Prior works using deep CNN for estimation of 3D pose from RGB image can be divided into two broad classes; viz. a) direct estimation of 3D joint locations [25, 46] and b) 2D heat-map projection followed by 3D pose estimation [28, 31, 26, 56]. To improve generalization on wild images,

adaptation from synthetic data [49, 49] or green-screen composition setups [28, 29] have been explored. Several approaches have successfully exploited structural [1, 29, 31] and temporal regularity [29, 56] as as an additional cue to further improve pose estimation performance [26]. Joint angle limit [1], bone length constraint [29] and relation between 3D interjoint distance and 2D inter-joint matrix [31] etc. are utilized as structural cues to enforce plausible 3D pose estimation. Some approaches [29, 56] utilize temporal smoothness in estimated 3D pose for consecutive video frames as an additional information to regularize sequential 3D pose estimation.

3. Approach

In this section we discuss different components which are proposed to effectively learn a generalized motion representation targeting unconstrained in-the-wild videos.

3.1. View-invariant pose embedding

The core objective behind separately learning a pose embedding representation [18] is to avoid enforcement of structural constraints in the subsequent stages of the motion modeling pipeline in a view-invariant setting. Moreover, while generating 3D pose, the model should completely restrain from generating unrealistic pose patterns.

View-invariant skeleton representation. Formally, let the pose in Global coordinate system (G) for a skeleton sample x_t^G be, $v_g^j(t) = [a_t^j, b_t^j, c_t^j]^T$, where $j \in (1, .., J)$ with J being the total number of skeleton joints. A canonical direction, v_t^n is defined to uniquely represent view-invariant pose coordinates, $v_c^j(t)$. v_t^n is obtained as the vector perpendicular to the XY-projection of the line segment joining the left-hip and right-hip point of $v_g^j(t)$. Following this, the global translation, $d_t^g = [a_t^g, b_t^g, c_t^g]^T$ and azimuth view-angle, γ_t^g (angle between v_t^n and the positive X-axis) is separated to form a new set of joints $v_c^j(t) = \mathcal{R}_{\gamma_t^g} \times (v_g^j(t) - d_t^g)$, with $\mathcal{R}_{\gamma_t^g}$ being the corresponding rotation matrix. Here d_t^g is the coordinate of the pelvis point in G. To achieve scale-invariance, we then normalize the bone lengths to that of a chosen canonical skeleton, and convert v_c^j to local parent coordinates for selected end-limb joints according to the kinematic structure of human skeleton. We follow Gram-Schmidt based approach inline with Akhter *et al*. [1] to obtain the pose representation for limb joint locations in the Local Coordinate System, L as, $v_l^j(t)$. We denote it as, x_t (i.e. $v_l^j(t)$ for $j = 1, ..J$), which effectively disentangles difficulties of modeling global position, camera-view and joint-location with respect to a single pelvis root.

Training of pose embedding network. We define encoder E_{pose}, decoder D_{pose} and discriminator $Disc_{pose}$ following similar architecture inspired from the kinematic tree of limb connections [5]. The inference network, E_{pose} takes the previously defined x_t as input representation as shown in Fig. 2A. Motivated from adversarial auto-encoder framework [27], to effectively leverage both continuity of pose embedding along with reduced reconstruction error, we employ an adversarial learning strategy in line with [19]. We denote the final pose embedding as $z_t = E_{pose}(x_t)$.

3.2. Data augmentation and pre-training

To simulate random unconstrained video augementations from static-camera feed, we perform smooth scale and focus variations with acceptable random parameter setting assuming orthographic projection. Besides this, time variations are also incorporated with fast and slow camera movements to simulate diverse unconstrained patterns as found in in-the-wild videos. By utilizing the output of a full-body tracker and 2D joint projection annotations (if available), we define a 1D sequence $\alpha_{0:T}$ capturing the proportion of visible body-part in the time interval 1 to T. Following this a temporal binary masking sequence is computed as $m_{0:T} = \alpha_{0:T} > 0.7$. However for samples with 2D pose annotation, we define a binary mask of J dimensions $m_{0:T}^J$ indicating visibility of individual joints.

Dataset selection. Here we discuss various datasets used in the proposed method to clearly explain the further training pipeline. Firstly, we consider two datasets with ground-truth 3D pose annotations i.e. Human 3.6M [10] and MPI-INF-3DHP [28], which is represented as $\mathcal{D}_{sup.}^{unsim.} = \mathcal{D}_{H3.6} \cup \mathcal{D}_{3DHP}$. Here *unsim.* stands for un-simulated frames without the videogramphic augmentations. Secondly, we choose two datasets without ground-truth 3D pose annotations i.e. MADS [55] and an in-house collection of 4 hours YouTube videos, which is represented as $\mathcal{D}_{unsup.}^{unsim.} = \mathcal{D}_{MADS} \cup \mathcal{D}_{YTube}$. The corresponding unconstrained simulated data is represented as $\mathcal{D}_{sup.}^{sim.}$ and $\mathcal{D}_{unsup.}^{sim.}$ respectively (see Table 1). We use the standard test-split of MADS dataset with and without random videographic augmentations, i.e. $\mathcal{D}_{test}^{unsim.}$ and $\mathcal{D}_{test}^{sim.}$ respectively as the test data for evaluation as it comes with the corresponding 3D pose ground-truth. Note that all these chosen datasets contains diverse disjoint motion or action categories to effectively highlight generalizability of the learned representation to unseen motion types.

Pretraining of CNN on $\mathcal{D}_{sup.}^{unsim.}$ We start from a Resnet-50 based CNN with pretrained parameters from Vnect [29]. The last layers after the *res-4f* block are fine-tuned to di-

Table 1. Dataset notations and their usage with supervision (sup.).

Settings	Train				Test	
	$\mathcal{D}_{H3.6}$	$\mathcal{D}_{3DHP}$	$\mathcal{D}_{MADS}$	$\mathcal{D}_{YTube}$	$\mathcal{D}_{MADS}^{test}$	$\mathcal{D}_{YTube}^{test}$
Pose sup.	✓	✓	×	×	-	-
w/o sim.	$\mathcal{D}_{sup.}^{unsim.}$		$\mathcal{D}_{unsup.}^{unsim.}$		$\mathcal{D}_{test}^{unsim.}$	$\mathcal{D}_{YTube}^{unsim.}$
with sim.	$\mathcal{D}_{sup.}^{sim.}$		$\mathcal{D}_{unsup.}^{sim.}$		$\mathcal{D}_{test}^{sim.}$	$\mathcal{D}_{YTube}^{sim.}$
Usage	for pre-training		For self-supervision		For testing	

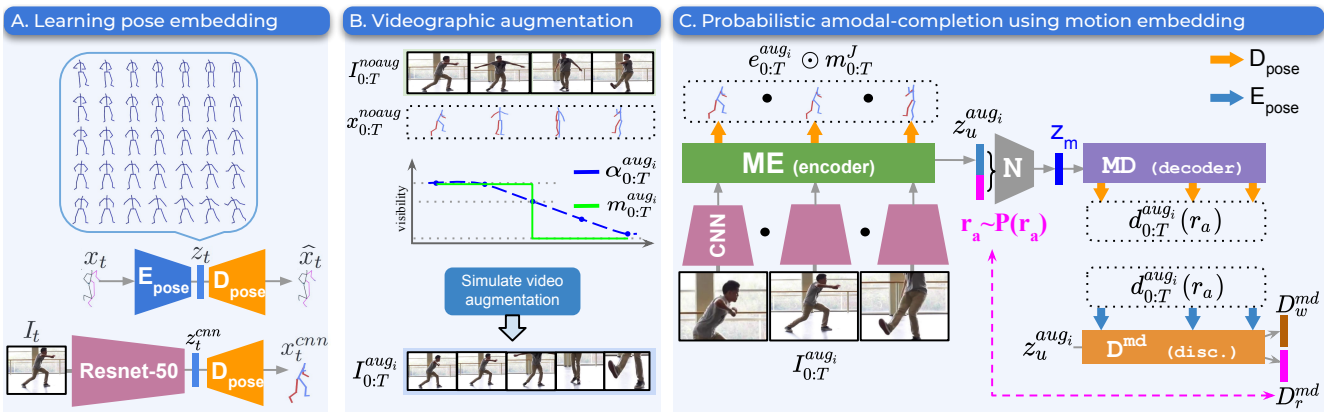

Figure 2. Overview-A) Architecture of pose embedding (Section 3.1) and CNN for pose estimation on the learned embedding space (Section 3.1). Grid interpolation of the learned continuous embedding space is also illustrated alongside. B) An illustration of $\alpha_{0:T}^{aug_i}$ and $m_{0:T}^{aug_i}$ as defined in Section 3.2 to simulate the unconstrained scenarios. C) Architecture of probabilistic motion embedding (Section 3.3).

Table 2. A list of important notations used in this paper.

Notations	Explanation (here RF denotes random factor)
$I_{0:T}^{noaug}$	The original unsimulated video
$I_{0:T}^{aug_k}$	kth random videographic augmentation
$x_{0:T}^{noaug}$	Ground-truth 3D pose of $I_{0:T}^{noaug}$
$e_{0:T}^{noaug}$	Pose prediction for $I_{0:T}^{noaug}$ at the output of *ME*
r_a	RF modeling uncertainty in amodal completion
$z_u^{aug_i}$	Uncertain motion embedding for ith simulation
z_m	Certain motion embedding; $z_m = N(z_u^{aug_i}, r_a)$
$d_{0:T}^{aug_i}(r_a)$	Pose prediction for $I_{0:T}^{aug_i}$ with RF r_a
$\hat{r}_a$	Predicted RF at the output of discriminator D_r^{md}
$\hat{r}_a^{aug_i}$	RF required to get an estimate of $x_{0:T}^{noaug}$ at the output MD as $d_{0:T}^{aug_i}(\hat{r}_a^{aug_i})$

rectly regress the latent *pose-vec* representation, denoted as z_t^{cnn} for a full-body input image I_t. However, instead of having a direct loss on z_t^{cnn}, we enforce a loss on $x_t^{cnn} = D_{pose}(z_t^{cnn})$ against x_t (see Fig. 2A) with frozen parameters of all the networks involved in the pose embedding framework (i.e. D_{pose} and E_{pose}). We denote this loss function as $\mathcal{L}(x_t^{cnn}, x_t) = |x_t^{cnn} - x_t|$. This is observed to improve performance considering the fact that, a direct loss on z_t enforces different weightage to the *pose-vec* depending on whether it is lying at a high-density or low-density region in the learned embedding space.

3.3. Learning temporal motion embedding

Aiming towards learning a general purpose motion representation, we plan to encode motion of short video snippets (short enough to be independent of action related cues). Before elaborating the individual modules, we provide a list of important notations with short descriptions in Table 2.

3.3.1 Probabilistic amodal motion completion

To leverage the temporal regularity from short video-clips (3 seconds), we introduce two bidirectional RNNs (Bi-LSTM) denoted as *ME* and *MD*. As shown in Fig. 2B, *ME* takes a sequence of *CNN* features denoted as $f_{0:T}^{CNN} = CNN(I_{0:T}^{aug_k})$ as the input representation. And it outputs a sequence of 3D pose denoted as $e_{0:T}^{aug_k}$ or $e_{0:T}^{noaug}$ (for kth simulation and un-simulated video respectively), through the frozen D_{pose} (shown as orange arrows in Fig. 2).

To effectively address the uncertainty in prediction of pose representation for frames with partial-visibility, we plan to model it as a conditional distribution over a combined representation of past and future motion, i.e. a transformation of the final bidirectional hidden representation of *ME* denoted as $z_u^{aug_i}$ (uncertain motion embedding for ith simulation). With z_m being the corresponding certain motion embedding (see Fig. 2B), we aim to model the randomness in $P(z_m|z_u^{aug_i})$ by introducing a random vector r_a drawn from fixed prior $P(r_a)$, i.e. $z_m = N(z_u^{aug_i}, r_a)$. Here, *MD* is designed as a bidirectional recurrent decoder, which decodes back the 3D pose sequence denoted as $d_{0:T}^{aug_i}(r_a)$ (through the frozen D_{pose}) from the certain motion embedding z_m obtained using a randomly drawn r_a.

Motivated by the training approach incorporated by BiHMP-GAN [18], we define a conditional discriminator with two output heads, D_r^{md} and D_w^{md} as shown in Fig. 2B. Conditioned on the given uncertain motion vector $z_u^{aug_i}$, the discriminator manages two separate tasks. Firstly, the output logit $D_w^{md}(d_{0:T}^{aug_i}(r_a); z_u^{aug_i})$ enforces the minimization of Wasserstein distance [8] with respect to the true distribution $x_{0:T}^{noaug}$ for the same given $z_u^{aug_i}$ conditioning. Secondly, $D_r^{md}(d_{0:T}^{aug_i}(r_a); z_u^{aug_i})$ retrieves the particular $\hat{r}_a$ vector which is used earlier to predict $d_{0:T}^{aug_i}(\hat{r}_a) = MD(N(z_u^{aug_i}, \hat{r}_a))$. Such a configuration also provides a new direction to incorporate a direct content loss $\mathcal{L}(d_{0:T}^{aug_i}(r_a^{aug_i}), x_{0:T}^{noaug})$; $\hat{r}_a^{aug_i} = D_r^{md}(x_{0:T}^{noaug}; z_u^{aug_i})$ even for the partially-visible frames without the use of masking (see Algo. 1). Here, $\hat{r}_a^{aug_i}$ is the random factor, which certainly outputs an estimate of $x_{0:T}^{noaug}$ as one of the plausible outcomes for the simulated input video-clip $I_{0:T}^{aug_i}$.

Training phase-1: The parameters of *ME*, *MD* and *N* are first trained using samples from both $\mathcal{D}_{sup.}^{unsim.}$ and $\mathcal{D}_{sup.}^{sim.}$ by enforcing $\mathcal{L}$ only for the visible time-steps i.e. $\mathcal{L}_A^{sup.} = \mathcal{L}(e_{0:T}^{aug_i}, x_{0:T}^{noaug}) \odot m_{0:T}^{J} + \mathcal{L}(d_{0:T}^{aug_i}(r_a), x_{0:T}^{noaug}) \odot m_{0:T}^{J}$

Training phase-2: After the first training phase, parameters of *N*, *MD* are finetuned along with the newly introduced discriminator D^{md} using an adversarial training framework shown in Algorithm 1.

3.4. Self-supervised learning

Here, we utilize the un-annotated data with and without videographic augmentation (i.e. $\mathcal{D}_{unsup.}^{sim.}$ and $\mathcal{D}_{unsup.}^{unsim.}$) to improve generalizability of the motion modeling pipeline for in-the-wild video samples. We formalize two probabilistic introspection based self-supervised techniques to form correspondence pairs of sequence of images which would have the same pose-sequence representation.

3.4.1 Self-supervision for frames with full visibility

Motivated by image based approaches relying on known synthetic warp-based augmentations [33], we plan to use diverse unconstrained camera simulations and temporal-shift [30] to gather correspondence pairs, which would have the same pose representation with full-body visibility. Suppose that, aug_i and aug_j are two different videography augmentations of the same video-clip, we impose correspondence only for the common frames with full-body visibility represented as the logical-and operation of the corresponding 1D visibility masks (see Fig. 3A), i.e.

$$\mathcal{L}_{A1}^{unsup.} = \mathcal{L}(d_{0:T}^{aug_i}(r_a), d_{0:T}^{aug_j}(r'_a)) \odot (m_{0:T}^{aug_i} \wedge m_{0:T}^{aug_j})$$

$\Theta_{MD}, \Theta_{D^{md}}$: Parameters of *MD* and D^{md}
for *k iterations* **do**
  for *m steps* **do**
    r_a: Sample random minibatch $\sim P(r_a)$
    $\mathcal{L}_{adv}^{disc} = D_w^{md}(d_{0:T}^{aug_i}(r_a); z_u^{aug_i}) - D_w(x_{0:T}^{noaug}; z_u^{aug_i})$
    $\mathcal{L}_{rec}^{r} = |r_a - \hat{r}_a|; \hat{r}_a = D_r^{md}(d_{0:T}^{aug_i}; z_u^{aug_i})$
    /* Parameter update for *Disc.* network*/
    $\Theta_{D^{md}} := \underset{\Theta_{D^{md}}}{\text{argmin}} \, (\mathcal{L}_{adv}^{disc} + \lambda_r \mathcal{L}_{rec}^{r})$
  end
  $\mathcal{L}_{content}^{X} = |x_{0:T}^{noaug} - d_{0:T}^{aug_i}(r_a^{aug_i})|$
  $\mathcal{L}_{adv}^{gen} = -D_w^{md}(d_{0:T}^{aug_i}(r_a);\ z_u^{aug_i})$
  /*Parameter Update for Decoder *MD*/*
  $\Theta_{MD} := \underset{\Theta_{MD}}{\text{argmin}} \, (\mathcal{L}_{adv}^{gen} + \lambda_r \mathcal{L}_{rec}^{r} + \lambda_c \mathcal{L}_{content}^{X})$
end

Algorithm 1: Training algorithm for the proposed probabilistic amodal limb completion.

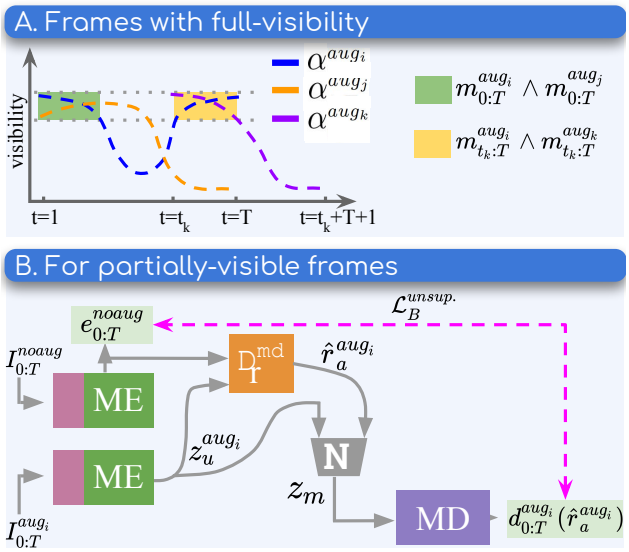

Figure 3. Schematic illustration of the self-supervised strategies using the pretrained probabilistic *MD* and *MF* (see Section 3.4).

Here, r_a and r'_a are randomly drawn from $P(r_a)$. Similarly, for the same video-clip we construct such correspondence pairs with a different augmentation of the temporally shifted sequence $I_{t_k:T+t_k}^{aug_k}$ (see Fig. 3A),

$$\mathcal{L}_{A2}^{unsup.} = \mathcal{L}(d_{t_k:T}^{aug_i}(r_a), d_{t_k:T}^{aug_k}(r'_a)) \odot (m_{t_k:T}^{aug_i} \wedge m_{t_k:T}^{aug_k})$$

We also configure the above two formulations for un-simulated video-clips by replacing $d_{0:T}^{aug_i}(r_a)$ with $d_{0:T}^{noaug}(r_a)$ and $m_{0:T}^{aug_i}$ with a vector of all ones. However, one must not construct such pairs for image sequences having partial-visibility keeping in mind the uncertainty in 3D pose estimate for such frames.

3.4.2 Self-supervision for partially-visible frames

Devoid of ignoring the partially-visible un-annotated video frames, we formalize a workaround by effectively utilizing the probabilistic motion embedding framework. Note that, along with the random unconstrained augmentations we also have access to the corresponding un-simulated full-body clips even for the un-annotated samples (i.e. $\mathcal{D}_{unsup.}^{unsim.}$). Let $e_{0:T}^{noaug}$ be the sequence of pose estimates for the un-simulated clip $I_{0:T}^{noaug}$ at the output of *ME*, and $d_{0:T}^{aug_i}(\hat{r}_a^{aug_i})$ is the sequence of pose estimates at the output of *MD* where $\hat{r}_a^{aug_i} = D_r^{md}(e_{0:T}^{noaug}; z_u^{aug_i})$ (see Fig. 3B) and $z_u^{aug_i} = ME(I_{0:T}^{aug_i})$. Hence forms a correspondence pair, i.e. $\mathcal{L}_B^{unsup.} = \mathcal{L}(d_{0:T}^{aug_i}(\hat{r}_a^{aug_i}), e_{0:T}^{noaug})$. Note that, this formulation is similar to $\mathcal{L}_{content}^{X}$ in Algo. 1.

Training phase-3 (self-supervision): We define, $\mathcal{L}_B^{sup.}$ by replacing e^{noaug} with the supervised ground-truth x^{noaug} in $\mathcal{L}_B^{unsup.}$. The final unsupervised and supervised loss functions are represented as, $\mathcal{L}^{unsup.} = \mathcal{L}_{A1}^{unsup.} + \mathcal{L}_{A1}^{unsup.} + \mathcal{L}_B^{unsup.}$ and $\mathcal{L}^{sup.} = \mathcal{L}_A^{sup.} + \mathcal{L}_{content}^{X} + \mathcal{L}_B^{sup.}$

respectively. Considering complexity of entanglement of multiple network components, we plan to adapt only the parameters of *ME* along with the last layer of the *CNN* for the un-annotated video samples. However, aiming to realize a unified framework across varied datasets and unconstrained videography setting, we incorporate a different training regime avoiding *mode-collapse* and other generalization issues. Alternate batches of supervised (i.e. $\mathcal{L}^{sup.}$) and unsupervised (i.e. $\mathcal{L}^{unsup.}$) training iteration is performed with separate Adam [15] optimizers.

4. Experiments

In this section, we present a thorough evaluation of our method for cross-dataset transfer of motion representation to unconstrained and unannotated video-clips. We provide a detailed ablation study of our self-supervised strategies along with comparisons to prior pose estimation methods. Further, we show qualitative results of probabilistic amodal limb completion.

Datasets Adhering to the detailed supervision strategies mentioned in Table 1 of Section 3.2, we use the simulated and un-simulated version of $\mathcal{D}^{test}_{MADS}$ for evaluation of the cross-dataset transfer performance. Whereas, the simulated and unsimulated version of the collected $\mathcal{D}_{YTube}$ dataset (unlabelled) is used for qualitative comparison. Table 3 depicts the diversity of motion classes among the selected datasets. It is to be noted that, the datasets based on laboratory setting have limited number of human subjects with fixed background scene. Therefore, keeping in mind the in-the-wild unconstrained scenarios found in internet videos, we ensured collection (i.e $\mathcal{D}_{YTube}$) of diverse human subjects (i.e. wearing diverse cloths in varied background) performing varied forms of motion (i.e. different dance forms such as western, modern, contemporary etc.). Maintaining a common evaluation ground against the prior arts we use the standard training split of Human3.6M [10], MPI-INF-3DHP [28] in our training iteration (i.e $\mathcal{L}^{sup.}$).

Implementation details We obtain human centered tight crops of size 224×224, using the ground truth 2D pose information for $\mathcal{D}^{sim.}_{sup.}$, $\mathcal{D}^{unsim.}_{sup.}$. However, for $\mathcal{D}^{sim.}_{unsup}$ and $\mathcal{D}^{unsim.}_{unsup}$, we obtain the detection bounding-box using Faster-RCNN [40]. Assuming it as a full-body bounding box we define the 1D $\alpha^{aug_i}_{0:T}$ and $m^{aug_i}_{0:T}$ for simulated samples in $\mathcal{D}^{sim.}_{unsup}$. Additionally, we normalize all videos to a common FPS of 30 with $T = 90$. We use single layer LSTM [3] with 512 hidden units for all the recurrent architectures shown in Fig. 2.

Table 3. Diversity of motion classes among the selected datasets

Dataset	Dynamic Motion Classes
$\mathcal{D}_H3.6$	Eating , Smoking, Discussion, Walking
$\mathcal{D}_YTube$	Indian, Modern, Contemporary etc.
$\mathcal{D}_MADS$	HipHop, Jazz, Sports, Taichi

4.1. Evaluation of the proposed approach

Evaluation metrics We evaluate ablations of our method against the previous arts on two metrics i.e. *PSS* and *MPJPE*. *PSS-Score* introduced in [16] assesses the full 3D pose as a structure, rather than independently looking at each joint. Hence, the metric is more robust as compared to conventional metric, *MPJPE*. Essentially, a scale-invariant performance score is obtained, by clustering the set of ground truth poses using k-means (using 50 clusters), and then computing cluster assignment accuracy. Maintaining a common ground, we perform view and bone-length normalization to the ground-truth pose as well as the view-variant pose predicted by the prior arts.

Our ablations We test three variants of our method, to validate different training phases. Our first baseline, is one where the training stops after phase-1 and phase-2, before employing any self-supervised strategy. We denote this baseline as *Ours-noSS*. Next, to validate the utility of self-supervision for frames with full-visibility, we include a baseline, where only $\mathcal{L}^{unsup}_{A1}$ and $\mathcal{L}^{unsup}_{A2}$ are enforced. We call this baseline as *Ours-SSA*. Finally, to validate the utility of the combined loss $\mathcal{L}^{unsup.}$, we train *Ours-SSAB*, where $\mathcal{L}^{unsup}_{B}$ is enforced in addition to $\mathcal{L}^{unsup}_{A1}$ and $\mathcal{L}^{unsup}_{A2}$.

Selection of prior arts Though there are several 3D pose estimation methods in literature, most of the approaches do not exploit the temporal video information [43, 16, 47]. Some recent papers such as [14, 39] showed improved performance on datasets such as Human3.6M by successfully leveraging the temporal cue. Some relatively earlier works such as VNect [29] employed post processing by imposing the temporal regularity. However, note that the above approaches do not handle pose estimation for partial visibility in an explicit manner as done by LCR++ [43] and Vosoughi *et al*. [51]. Contrastingly, the proposed method aims to leverage both temporal and partial-visibility in the process of developing a probabilistic amodal limb completion framework. Therefore, to have a fair comparison, we evaluate our approach against all the above discussed state-of-the-art methods. However, these methods have to be re-evaluated on our intended task (unsupervised adaptation), as they do not report numbers on it. We therefore use the publicly available implementation with pretrained weights provided by the authors for [43, 29, 16, 47, 39]. Additionally, although the code for [51, 14] has not been made public, owing to the simplicity of these methods, we reimplement them and reproduce their reported numbers.

4.1.1 Comparison of 3D pose estimation

As described in the Section 1, our primary goal of human motion representation learning has a lack of prior literature for us to perform a thorough quantitative compari-

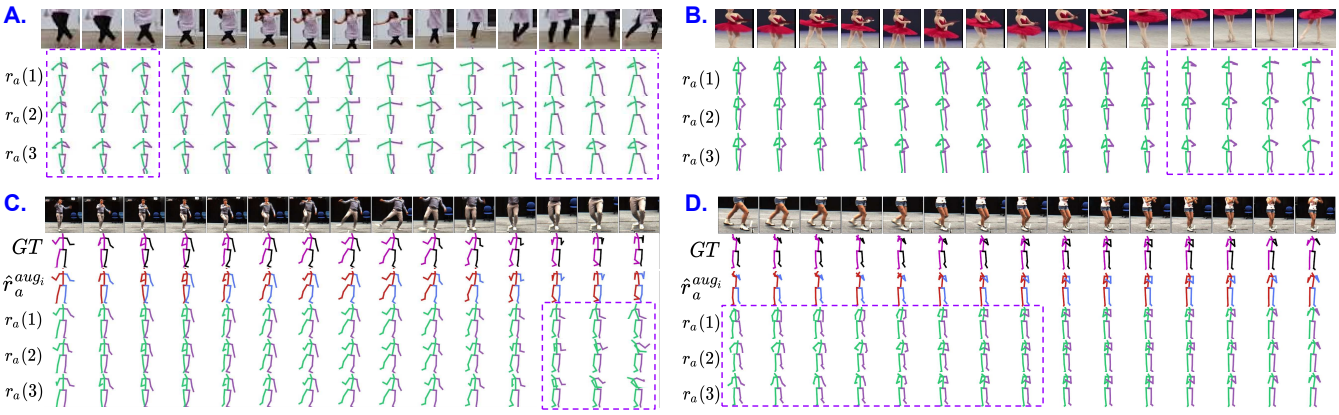

Figure 4. Qualitative results on probabilistic amodal limb completion on unconstrained in-the-wild dataset, D^{test}_{YTube}(top panel A and B), and D^{test}_{MADS} (bottom panel C and D). The purple dotted line indicates examples where diverse predictions are obtained only for the joints that are truncated in the given image. Here, $\hat{r}_a^{aug_i}$ indicates the r_a obtained by passing the ground truth pose sequence to the discriminator. Notice that, our framework is capable of generating multiple plausible pose-filling outcomes particularly for the non-visible body-joints, towards both beginning or end of the temporal-sequence as a result of the robust bi-directional motion representation.

Table 4. Comparison of cross-dataset 3D pose estimation results on $\mathcal{D}^{test}_{MADS}$. All methods have no access to 3D pose supervision from $\mathcal{D}^{test}_{MADS}$(i.e. "cross-dataset"). For PSS (Pose Structure Score) the higher (↑) the better, and for MPJPE the lower (↓) the better. MPJPE here is abbreviated as MP.

Method	Dance Style					
	HipHop	Jazz	Taichi	Sports	Avg	
	PSS(↑)	PSS(↑)	PSS(↑)	PSS(↑)	PSS(↑)	MP(↓)
Un-simulated Videos						
Vosoughi *et al.* [51]	67.4	73.9	62.5	30.3	58.5	1.14
LCR++ [43]	84.3	86.3	43.7	29.4	60.9	1.04
VNect [29]	69.2	67.5	74.5	34.5	61.4	1.13
Kocabas *et al.* [16]	74.3	72.9	66.8	39.3	63.3	1.01
Sun *et al.* [47]	77.8	78.1	69.3	40.1	66.3	0.93
Tekin *et al.* [14]	82.3	81.7	70.2	37.9	68.0	0.95
Pavllo *et al.* [39]	83.2	82.3	70.4	40.5	69.1	0.93
Ours-noSS	83.4	83.1	72.3	42.1	70.2	0.90
Ours-SSA	86.2	85.2	75.4	46.7	73.4	0.84
Ours-SSAB	**87.1**	**86.5**	**76.2**	**48.1**	**74.5**	**0.81**
Simulated Videos						
Sun *et al.* [47]	54.5	51.3	41.8	33.4	45.3	1.58
Kocabas *et al.* [16]	52.6	52.1	43.4	35.4	45.9	1.61
VNect [29]	54.5	54.8	60.1	26.6	49.0	1.26
LCR++ [43]	70.2	73.4	35.2	23.2	50.5	1.21
Vosoughi *et al.* [51]	66.4	64.2	58.9	32.1	55.4	1.19
Tekin *et al.* [14]	69.5	70.4	58.3	30.1	57.1	1.16
Pavllo *et al.* [39]	73.3	69.6	63.8	39.1	61.4	1.07
Ours-noSS	78.4	75.2	69.2	41.9	66.2	0.97
Ours-SSA	79.5	76.4	70.6	42.4	67.2	0.95
Ours-SSAB	**82.8**	**80.4**	**72.4**	**45.1**	**70.2**	**0.90**

son on this specific task. However, pose estimation performance on datasets not seen during the supervised learning phase (i.e cross-dataset transfer) can validate the utility of our self-supervision strategies. We therefore choose to evaluate pose estimation performance on $\mathcal{D}^{test}_{MADS}$, a dataset not used during the supervised learning phase. We follow two strategies for evaluation. For un-simulated videos, the pose estimates from e_t^{noaug} are compared against ground-truth pose. Whereas, for simulated video, the metrics are computed on the pose estimates from $d_t^{aug_i}(r_a)$. Considering

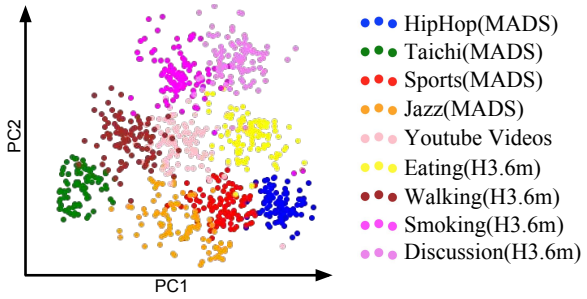

Figure 5. PCA plot generated using $\mathcal{D}^{test}_{MADS} \cup \mathcal{D}^{test}_{H3.6} \cup \mathcal{D}^{test}_{YTube}$ showing generalizability of the learned motion representation across datasets with diverse action categories. Additionally, the plot clearly shows that our model learns a common embedding space for multiple datasets.

the uncertainty in 3D pose as a result of partial-visibility, we follow the evaluation protocol described in [18], where *min-error* and *best-PSS* is reported over a batch of 500 randomly drawn, r_a vectors. However, for the prior-arts, we follow the protocol proposed by the respective works for both simulated and un-simulated inputs. Results in Table 4 clearly highlights our superiority in cross-dataset transfer against our baselines and prior-arts. More specifically, we see a substantial improvement on simulated videos as a result of the proposed self-supervised strategy.

We further observe that addition of self-supervised loss $\mathcal{L}_A^{unsup}$ leads to a small improvement in performance on simulated videos, but a significant improvement on unsimulated videos, due to the fact that this loss takes into account only the frames which have full visibility. In contrast, as self-supervised loss $\mathcal{L}_B^{unsup}$ enforces self-supervision for partially visible frames, we see that enforcing this loss leads to a small improvement in performance on un-simulated videos, but a significant improvement on simulated videos.

4.1.3 Qualitative results

In Fig. 4, we present qualitative results on probabilistic amodal limb completion. For this task, it can be observed here that our model successfully reconstructs the visible joints deterministically. At the same time, by varying the random vector, r_a we can generate diverse predictions with high variance for the regions that are not visible. Refer to the supplementary for a more quantitative validation of our probabilistic model in the form of variance curves. Additionally, a PCA plot is shown in Fig. 5, that depicts motion class separation in the learnt motion embedding space.

5. Conclusion

In this work, we present a novel method for cross-dataset transfer of human motion representation. A probabilistic framework is proposed which disentangles the uncertainty inherent in unconstrained natural videos, and enables the application of effective self-supervised losses. *State-of-the-art* results on tasks such as pose estimation, particularly on datasets not seen during training, substantiates our model's generalizability and robustness.

Acknowledgements. This project was partly supported by Indo-UK project (DST/INT/UK/P-179/2017), DST, India; Uchhatar Avishkar Yojana (UAY) project (IISC_010), MHRD, India; and a Wipro PhD Fellowship (Jogendra). We would like to thank Maharshi Gor for helpful discussions.

References

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1455, 2015.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[4] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *The IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[5] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015.

[6] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. Marker-less 3d human motion capture with monocular image sequence and height-maps. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[7] K. Ehsani, R. Mottaghi, and A. Farhadi. Segan: Segmenting and generating the invisible. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[9] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2014.

[11] D. Jayaraman and K. Grauman. Learning image representations tied to ego-motion. In *The IEEE International Conference on Computer Vision*, pages 1413–1421, 2015.

[12] A. Kanazawa, D. W. Jacobs, and M. Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2016.

[13] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Amodal completion and size constancy in natural scenes. In *The IEEE International Conference on Computer Vision*, pages 127–135, 2015.

[14] I. Katircioglu, B. Tekin, M. Salzmann, V. Lepetit, and P. Fua. Learning latent representations of 3d human pose with deep neural networks. *International Journal of Computer Vision*, pages 1–16, 2018.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019.

[17] J. N. Kundu, M. Gor, D. Agrawal, and R. V. Babu. Gan-tree: An incrementally learned hierarchical generative framework for multi-modal data distributions. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[18] J. N. Kundu, M. Gor, and R. V. Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *AAAI Conference on Artificial Intelligence*, 2019.

[19] J. N. Kundu, M. Gor, P. K. Uppala, and V. B. Radhakrishnan. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1467. IEEE, 2019.

[20] J. N. Kundu, P. Krishna Uppala, A. Pahuja, and R. Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] J. N. Kundu, N. Lakkakula, and R. V. Babu. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[22] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 840–849. IEEE, 2017.

[23] K. Li, S. Li, S. Oh, and Y. Fu. Videography-based unconstrained video analysis. *IEEE Transactions on Image Processing*, 26(5):2261–2273, 2017.

[24] K. Li and J. Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016.

[25] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *The IEEE International Conference on Computer Vision*, pages 2848–2856, 2015.

[26] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng. Recurrent 3d pose sequence machines. In *Computer Vision and Pattern Recognition, 2017 IEEE Conference on*, pages 5543–5552. IEEE, 2017.

[27] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[28] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision, 2017 International Conference on*, pages 506–516. IEEE, 2017.

[29] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4):44, 2017.

[30] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[31] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017.

[32] M. M. Murray, D. M. Foxe, D. C. Javitt, and J. J. Foxe. Setting boundaries: brain dynamics of modal and amodal illusory shape completion in humans. *Journal of Neuroscience*, 24(31):6898–6903, 2004.

[33] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[34] D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3637–3645, 2018.

[35] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, pages 801–816. Springer, 2016.

[36] D. Pathak, R. Girshick, P. Dollar, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.

[37] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[38] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017.

[39] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[40] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[41] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6148–6157, 2017.

[42] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, United States, 2017.

[43] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[44] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[45] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[46] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *2017 IEEE International Conference on Computer Vision*, pages 2621–2630. IEEE, 2017.

[47] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *The European Conference on Computer Vision*, September 2018.

[48] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, pages 844–855, 2017.

[49] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

[50] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[51] S. Vosoughi and M. A. Amer. Deep 3d human pose estimation under partial body presence. In *2018 25th IEEE International Conference on Image Processing*, pages 569–573. IEEE, 2018.

[52] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *The IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[53] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.

[54] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[55] W. Zhang, Z. Liu, L. Zhou, H. Leung, and A. B. Chan. Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3d human pose estimation. *Image and Vision Computing*, 61:22–39, 2017.

[56] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4966–4975, 2016.

[57] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.

[58] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollar. Semantic amodal segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017.