Non-local Latent Relation Distillation for Self-Adaptive 3D Human Pose Estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Available 3D human pose estimation approaches leverage different forms of strong (2D/3D pose) or weak (multi-view or depth) paired supervision. Barring synthetic or in-studio domains, acquiring such supervision for each new target environment is highly inconvenient. To this end, we cast 3D pose learning as a self-supervised adaptation problem that aims to transfer the task knowledge from a labeled source domain to a completely unpaired target. We propose to infer *image-to-pose* via two explicit mappings viz. *image-to-latent* and *latent-to-pose* where the latter is a pre-learned decoder obtained from a prior-enforcing generative adversarial auto-encoder. Next, we introduce relation distillation as a means to align the unpaired cross-modal samples *i.e*. the unpaired target videos and unpaired 3D pose sequences. To this end, we propose a new set of non-local relations in order to characterize long-range latent pose interactions unlike general contrastive relations where positive couplings are limited to a local neighborhood structure. Further, we provide an objective way to quantify non-localness in order to select the most effective relation set. We evaluate different self-adaptation settings and demonstrate state-of-the-art 3D human pose estimation performance on standard benchmarks.

1 Introduction

Human pose estimation systems have garnered immense attention due to their innumerable applications [51, 19, 80]. The successes of such systems are mostly driven by large-scale datasets containing images with paired 3D pose annotations [26]. Unlike 2D joint landmarks, annotating a 3D pose requires body-worn sensors or multi-camera structure-from-motion setup [66] which is challenging to install outdoors. Hence, the available datasets are either captured in constrained laboratory settings or are limited in size and diversity. Unsurprisingly, models trained on such datasets exhibit poor cross-dataset generalization. Addressing this, several approaches [11, 72, 52, 34] resort to weakly-supervised methods that rely on images with paired 2D landmark annotations. Certain methods require additional supervision such as depth [76, 11] or multi-view image pairs [61, 33]. However, they still suffer from dataset-bias due to their strong reliance on some form of paired supervision.

In this work, we intend to digress from any form of paired supervision or auxiliary cues (multi-view or depth) thereby avoiding the curse of dataset-bias. To this end, we introduce a self-supervised domain adaptation framework for 3D human pose estimation. In the proposed setting, we consider access to three different datasets. First, a labeled source dataset obtained from graphics-based synthetic environments such as SURREAL [74] or in-studio datasets such as Human3.6M [26]. Second, unlabeled video sequences from the target domain. Third, a set of unpaired 3D pose sequences. Following this, *image-to-pose* inference is carried out via two explicit mappings *i.e*., *image-to-latent* CNN followed by a *latent-to-pose* network. Here, the *latent-to-pose* network is pre-learned as a decoder of a prior-enforcing generative adversarial auto-encoder [43] (AAE) in order to restrict the pose predictions within the plausibility limits. We follow the same to pre-learn a generative motion embedding via a recurrent AAE setup which operates on the latent pose space instead of the raw pose

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

sequences. Both pose and motion embeddings are trained solely on the unpaired 3D pose dataset. Next, we prepare *image-to-latent* CNN by supervising on the labeled source dataset. After finishing the above pre-learning steps, our prime objective is to train or adapt only the *image-to-latent* CNN so that it can perform well on the unlabeled target domain samples. In order to align the output manifold of the *image-to-latent* with the pre-learned latent pose manifold (in the absence of cross-modal pairs), one must formalize innovative ways to represent the *dark-knowledge* based on which the student (*i.e*. *image-to-latent*) output can align to that of the teacher (*i.e*. *pose-to-latent*).

Several studies [18, 20] on human cognitive development advocate that, in a self-supervised paradigm, new knowledge is acquired by relating entities based on some semantic rule. Motivated by this, we explore different ways of formalizing inter-entity relations. A relation can be characterized as lower order or higher order based on the number of entities that participate in defining a relationship tuple. For instance, in contrastive learning [53, 12], a pose space triplet relation expresses a coupling of just 3 pose entities, and is thus considered a lower order relation. However, a similar contrastive triplet defined in the hierarchical motion space (*i.e*. temporal pose sequences of fixed sequence length T) would couple $3T$ pose entities, thus is considered a higher order relation.

Next, we adapt the target-specific *image-to-latent* by minimizing the relational energies derived from the contrastive relations. Unlike in prior-arts [12, 47] where the output embedding is learned from scratch, we are restricted to operate on a pre-learned output pose embedding. Consequently, the model often converges to a degenerate solution exhibiting instance-level misalignment [42]. We realize that the positive coupling in contrastive relations is limited to local neighborhood structures resulting in sub-optimal alignment. This motivates us to develop a new set of relations that would express positive coupling of diverse non-local relations (beyond the structural neighborhood), thereby characterizing long-range latent pose interactions in a much effective manner.

We define tangible non-local relations separately in the pose and motion space that are categorized under a) lower order non-local relations and b) higher order non-local relations, respectively. For instance, "`pose-flip`" is a lower order non-local relation which associates anchor poses with their left-right flipped versions. Similarly, "`flip-backward`" is a higher order non-local relation which couples anchor pose-sequences with their flip-backward counterpart which is obtained via temporal reversal (backwards) of the individually flipped frames. Here, the corresponding relational energy is devised via latent space relation transformers. These are essentially frozen neural networks that are trained to regress the latent embedding of the relational counterpart given latent embedding of the anchor as input. In a nutshell, the relational energies aim to preserve the equivariance of higher order spatio-temporal relations between the two modalities as a means to perform the cross-modal alignment. It turns out that, among various relations, relations coupling the most diverse non-local samples result in a better cross-modal alignment. We perform extensive experiments to validate the efficacy of our approach and demonstrate superior generalizability on samples from in-the-wild environments. We summarize our contributions as follows:

- The proposed solution for self-adapting 3D human pose model involves cross-modal alignment between the unpaired samples from the input and output modalities, via relation distillation. Highly non-local instances are associated using novel relation transformer networks to specifically cater to the instance-level misalignment.
- We provide insights to select the most effective non-local relations. This involves quantifying non-localness of a relation as the average latent-distance between the coupled entities.
- We evaluate different self-adaptation settings and demonstrate state-of-the-art 3D human pose estimation performance against the available semi-supervised and weakly-supervised prior arts on Human3.6M [26], MPI-INF-3DHP [45], and 3DPW [75].

2 Related Works

Table 1 shows a comparison of the proposed setting against the prior-arts. Our framework is equipped with self-adaptation capability while utilizing unpaired 3D pose and synthetic 3D pose supervision.

Knowledge distillation (KD). In KD [24] the student model focuses on mimicking the output behaviour of a pre-trained teacher model. Traditional techniques aim to match instance level behaviours such as matching the output activation [63], gram matrix [82], gradient [67], attention maps [83], etc. Recently, several techniques proposed to align pairwise distance [71, 58] or similarity graphs [54, 39] in an effort to look beyond the instance-level matching. Further, to capture higher order output dependencies, researchers adopt techniques [70] inspired from the advances in self-supervised contrastive

learning literature [53, 4]. Essentially, a balanced distillation of both structural and instance-level knowledge leads to superior generalization [78]. However, in the proposed setting, the teacher, *pose-to-latent* and the student, *image-to-latent* do not share the same input modality. Further, the absence of paired correspondence restricts us from directly employing such distillation techniques.

Monocular 3D human pose estimation. There are two broad categories of models that infer 3D poses from a single image, *i.e.* a) those using model-based parametric representation [5, 8, 57, 7] and b) others that infer the 3D pose representation directly [64, 2, 6]. Models belonging to the first category map the input image to the latent parameters of a pre-defined parametric human model. This setting establishes the base to impose various pose priors through techniques such as adversarial training. Those belonging to the second directly map the input image to the corresponding 3D pose. These models are further categorized into a) one-stage methods [88, 68, 55, 50, 56, 81] that directly map input image to the 3D pose, and b) two-stage methods [87, 44, 49, 25] that adopt a mapping from the image to an intermediate 2D pose, followed by lifting of 2D-to-3D. In our work, the shared latent pose can be seen as a parametric form to represent plausible 3D poses.

Table 1: Characteristic table comparing the positive and negative attributes of ours against the prior works.

Methods	Real Paired Sup.			Unpaired	Synthetic	Self-
	3D pose	2D pose	Multi view	3D pose Sup.	3D pose Sup.	adaptation capability
Rhodin *et al*. [60]	✓	✗	✓	✗	✗	✗
Chen *et al*. [13]	✗	✓	✓	✗	✗	✗
Wandt *et al*. [76]	✗	✓	✗	✓	✗	✗
Chen *et al*. [11]	✗	✓	✗	✓	✗	✗
Doersch *et al*. [16]	✗	✓	✗	✗	✓	✗
Iqbal *et al*. [27]	✗	✓	✓	✗	✗	✗
Kundu *et al*. [35]	✗	✓	✗	✓	✗	✗
Zhang *et al*. [85]	✗	✓	✗	✗	✗	✓
Ours	✗	✗	✗	✓	✓	✓

Domain Adaptation. A number of works aim to tackle the problem of domain shift induced due to the vastly different training and deployment scenarios. For the animal pose estimation task, Cao *et al*. [9] apply discriminator based discrepancy minimization technique. Zhang *et al*. [86] leverage multi-modal input such as depth and body segmentation masks. Similarly, Doersch *et al*. [16] utilize optical-flow and 2D key-points as input representations. Further, Zhang *et al*. [85] propose to use 2D landmarks, obtained from a powerful off-the-shelf 2D pose detector, as the input modality. They perform instance-level geometry-aware self supervised learning on each target for inference time 2D-to-3D lifting. These approaches rely on alternative input modalities that are least affected by domain shift unlike the raw RGB images. The proposed approach does not utilize any such auxiliary inputs thus addressing one of the most challenging adaptation settings.

3 Approach

In the following, we first define the basic notations and network components required to discuss the proposed training setup. Then we provide more details about the proposed self-supervised adaptation.

3.1 Notations and pre-learning steps

We start with listing the available datasets based on which we introduce our learning setup.

3.1.1 Datasets. We consider access to three different datasets as follows (refer Fig. 1). **a)** A labeled source dataset $(x^s, y^s) \in \mathcal{D}^s$, where (x^s, y^s) denotes a tuple of a source image paired with its 3D pose annotation, **b)** A set of unlabeled video sequences from the target domain, $X = [x_1, x_2, .., x_T] \in \widetilde{\mathcal{X}}$ where T denotes the sequence length. Here, $x_t \in \mathcal{X}$ denotes a single RGB image frame sampled from the target domain sequence X. And, **c)** a set of unpaired 3D pose sequences, $Y = [y_1, y_2, .., y_T] \in \widetilde{\mathcal{Y}}$. Here, a single 3D pose frame is represented as $y_t \in \mathcal{Y}$.

3.1.2 Network components. The *image-to-pose* inference is to be carried out via the *image-to-latent* CNN, $\mathtt{G} : \mathcal{X} \rightarrow \mathcal{Z}$ followed by a *latent-to-pose* network, $\mathtt{D_p} : \mathcal{Z} \rightarrow \mathcal{Y}$ (see Fig. 1C). We introduce two latent embeddings, i.e., pose embedding and motion embedding, as follows.

a) The pose embedding is denoted as $z \in \mathcal{Z}$. The latent space $\mathcal{Z}$ is constrained to follow a uniform prior distribution, $z \in [-1, 1]^{32}$. This is realized by training an AAE (adversarial auto-encoder [43]) whose encoder and decoder mappings are represented as, $\mathtt{E_p} : \mathcal{Y} \rightarrow \mathcal{Z}$ and $\mathtt{D_p} : \mathcal{Z} \rightarrow \mathcal{Y}$ (Fig. 1A).

b) The motion embedding is denoted as $v \in \mathcal{V}$. In line with the preparation of pose embedding, the encoder and decoder mappings of the motion-AAE (see Fig. 1B) are denoted as $\mathtt{E_m} : \widetilde{\mathcal{Z}} \rightarrow \mathcal{V}$ and $\mathtt{D_m} : \mathcal{V} \rightarrow \widetilde{\mathcal{Z}}$. Here, the motion embedding is learned as a hierarchical temporal embedding

of sequence of pose embeddings, $Z = [z_1, z_2, .., z_T] \in \widetilde{\mathcal{Z}}$. Moreover, the recurrent auto-encoder operates on the sequence of pose embeddings ($Z \in \widetilde{\mathcal{Z}}$) instead of the raw pose sequences ($Y \in \widetilde{\mathcal{Y}}$).

3.1.3 Pre-learning steps. We introduce separate domain-specific *image-to-latent* mappings for the source and target domains *i.e.*, $\mathtt{G^s} : \mathcal{X}^s \rightarrow \mathcal{Z}$ and $\mathtt{G} : \mathcal{X} \rightarrow \mathcal{Z}$ respectively. The pose and motion auto-encoders are trained using the unpaired 3D pose data, $Y \in \widetilde{\mathcal{Y}}$ and kept frozen in later training stages (see Fig. 1). Here, $\mathtt{G^s}$ is prepared by training on $\mathcal{D}^s$ using the following loss term: $||\mathtt{D_p} \circ \mathtt{G^s}(x^s) - y^s||$, where $\circ$ denotes functional composition. Given a target image-sequence (or video) X, its motion embedding is obtained as $\hat{v} = \mathtt{E_m} \circ \mathtt{G}(X)$. Here, $\mathtt{G}(X)$ separately processes each temporal frame of X to obtain the latent pose sequence $\hat{Z}$ which is passed through $\mathtt{E_m}$ to obtain $\hat{v}$. Sec. 3.3 details the procedure to adapt $\mathtt{G}$ after initializing it from $\mathtt{G^s}$.

3.2 Problem formulation

The prime objective is to train or adapt the *image-to-latent* mapping, $\mathtt{G}$ to better generalize it to the unlabeled target domain samples (*i.e.* x or X).

a) Domain adaptation perspective. One can see it as an unsupervised domain adaptation (DA) problem [17, 65] which aims to adapt a source trained $\mathtt{G^s}$ in order to obtain a target specific mapping $\mathtt{G}$. However, unlike segmentation or classification, 3D pose is a structured regression output. Consequently, the usual statistical or adversarial discrepancy minimization based [73, 40] DA solutions usually fail to improve the adaptation performance against the corresponding pre-adaptation baselines [16].

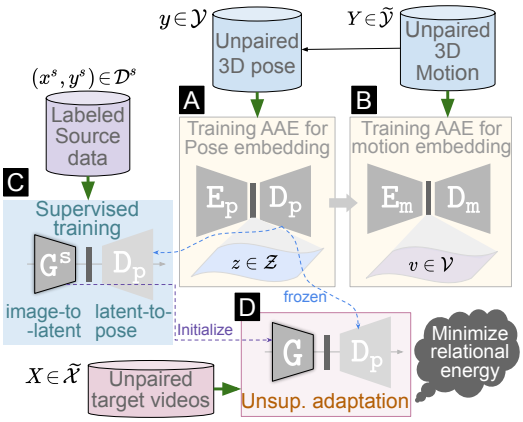

Figure 1: Proposed training setup. **A** and **B**. Training pose and motion AAEs on the unpaired pose and motion data. **C.** Initializing $\mathtt{G}^s$ by supervising on labeled source data. **D.** Adapting $\mathtt{G}$ on unlabeled target videos.

b) Knowledge distillation perspective. The problem in hand can also be perceived as a manifold alignment problem [77] where the output manifold of the *image-to-latent* $\mathtt{G}$, needs to be aligned with the pre-learned pose manifold, $\mathcal{Z}$. In knowledge distillation (KD), the same is achieved by formalizing the *dark-knowledge* based on which the output of a student network can align to that of a teacher network. However, in the proposed setting, the teacher, *pose-to-latent* $\mathtt{E_p} : \mathcal{Y} \rightarrow \mathcal{Z}$ and the student, *image-to-latent* $\mathtt{G} : \mathcal{X} \rightarrow \mathcal{Z}$ do not share the same input modality (*i.e.* image space $\mathcal{X}$ vs. pose space $\mathcal{Y}$). The most common form of distillation aims to align the instance-wise embeddings [24, 63]. However, such approaches are infeasible in the absence of cross-modal pairing *i.e.*, $(x, y) \in \mathcal{X} \times \mathcal{Y}$ in the unlabeled target domain. Though one can obtain pseudo (x, y) pairs by forwarding the unlabeled target images through the source specific $\mathtt{G^s}$ (via $\mathtt{D_p} \circ \mathtt{G^s} : \mathcal{X} \rightarrow \mathcal{Y}$), the obtained pose predictions are highly unreliable due to the input domain-shift. This motivates us to look for innovative ways to represent the *dark-knowledge* without relying on cross-modal pairs.

3.3 Proposed solution

We propose to represent the *dark-knowledge* by forming relational associations that aim to relate a group of intra-modal entities. For example, distilling triplet associations as used in self-supervised contrastive learning [12, 47] literature can be seen as a trivial form of representing the *dark-knowledge*.

3.3.1 Distilling local neighborhood relations via contrastive learning

General contrastive learning can be applied at both latent pose space, $\mathcal{Z}$ and motion space, $\mathcal{V}$.

a) Lower-order contrastive. Pose space contrastive learning relies on relational associations of the form $\mathcal{C}^{\mathtt{x}} = \{(x_t, x_t^+, \{x_t^-\})$ where $(x_t, x_t^+) \rightarrow$ *positive* and $(x_t, \{x_t^-\}) \rightarrow$ *negative*$\}$ (Fig. 2C, bottom panel). Here, $\{x_t^-\}$ represents a batch of unique instances that are in no way related to x_t. *positive* indicates a positive coupling of x_t and x_t^+. Here, x_t^+ is obtained via a pose-invariant image space augmentation of x_t *i.e.*, $x_t^+ = \mathtt{A}(x_t)$, and $\mathtt{G}(x_t) = \mathtt{G}(x_t^+)$. However, in the *negative* coupling, one just knows that the corresponding poses are unrelated. Thus, $\mathcal{C}^x$ mostly focuses on characterizing

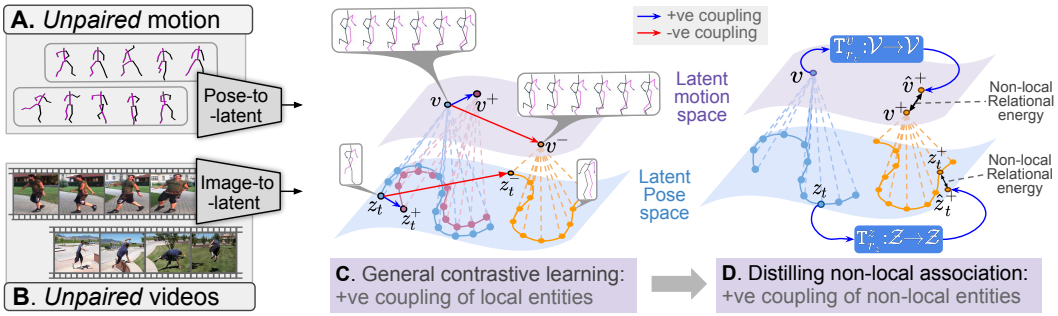

Figure 2: We align samples from unpaired pose (or motion) and unpaired images (or videos) at a shared latent pose space by distilling higher order (associating multiple instance via motion) non-local (*e.g. flip-backward*) relations. Relations are equivalent to a form data-interlinking as done in knowledge-graphs.

the lower-order latent local neighborhood structure. Drawing inspiration from InfoNCE [53], the corresponding relational energy loss (with τ as the temperature hyperparameter) is represented as,

$$\mathcal{L}_{LCR} = \sum_{(x_t, x_t^+, \{x_t^-\}) \in \mathcal{C}^x} e^{(G(x_t).G(x_t^+)/\tau)} / (e^{(G(x_t).G(x_t^+)/\tau)} + \sum_{\{x_t^-\}} e^{(G(x_t).G(x_t^-)/\tau)}) \quad (1)$$

b) Higher-order contrastive. Similarly, motion space contrastive characterizes higher-order local neighborhood relation *i.e.*, $\mathcal{C}^{\mathbf{X}} = \{(X, X^+, \{X^-\})$ where $(X, X^+) \rightarrow$ *positive* and $(X, \{X^-\}) \rightarrow$ *negative*$\}$. The corresponding relational energy is denoted as $\mathcal{L}_{HCR}$. $\mathcal{L}_{HCR}$ takes the same form as of Eq 1 by replacing $\mathtt{G}(x_t)$ with $\mathtt{E_m} \circ \mathtt{G}(X)$, and similarly for X^+ and X^- (Fig. 2C, top panel).

Higher-order vs lower-order relations. A natural question that arises is: *why to use motion embedding when the goal task is to realize an image-to-pose mapping?*. We hypothesize that formalization of higher-order pose association would enhance the expressibility of the *dark-knowledge* which is of prime interest towards realizing superior cross-modal alignment. Here, higher-order refers to relational association of a large number of pose entities. In $\mathcal{C}^X$, each individual triplet relation expresses a positive coupling of $2T$ pose entities against just 2 in $\mathcal{C}^x$. Thus, the hierarchical motion embedding facilitates a suitable platform to formalize temporally structured (temporal order must be retained) higher-order relational associations of the pose space entities (see Fig. 2C).

3.3.2 Distilling non-local relations via equivariance consistency

We observe that adapting the *image-to-latent* mapper $\mathtt{G}$ just by minimizing $\mathcal{L}_{LCR}$ and $\mathcal{L}_{HCR}$ results in poor pose estimation performance. In general self-supervised contrastive learning prior-arts, the latent embedding is learned from scratch alongside the training of *image-to-latent* mapping. However, in the proposed setting, we aim to adapt the *image-to-latent* network $\mathtt{G}$ in order to align its output with the pre-learned pose and motion embeddings (*i.e.* the latent embeddings of frozen $\{\mathtt{E_p}, \mathtt{D_p}\}$ and $\{\mathtt{E_m}, \mathtt{D_m}\}$). Consequently, we need to address a special degenerate scenario where the distillation losses completely converge even when the model exhibits severe instance-level misalignment [42].

Local vs non-local relations. We hypothesize that relations having non-local positive coupling, equivalent to relations in a knowledge-graph (KG), are the best way to express the instance-grounded *dark-knowledge*. Note that, the positive coupling in contrastive-based relations are limited to local neighborhood, as here, the positive counterpart of an anchor is obtained via simple pose-invariant augmentations (*i.e.* $x_t^+ = \mathtt{A}(x_t)$ in both $\mathcal{C}^x$ and $\mathcal{C}^X$). To this end, our goal is to come up with a new set of relational associations with non-local positive couplings. This aims to characterize long-range latent pose interactions, *i.e.* a positive coupling of two entities that are grounded far away in the latent pose or motion space (see Fig. 3C). To this end, we introduce relation transformers.

Relation transformers. Relation transformers operating on the pose and motion embeddings are neural network mappings of the form, $\mathtt{T}^z_{r_z} : \mathcal{Z} \rightarrow \mathcal{Z}$ and $\mathtt{T}^v_{r_v} : \mathcal{V} \rightarrow \mathcal{V}$ respectively. Here, r_z and r_v denote independent embedding specific rule-ids for the pose and motion embeddings respectively. The equivalent rule-specific image-transformers operating on the image and image-sequence are represented as $\mathtt{T}^x_{r_z} : \mathcal{X} \rightarrow \mathcal{X}$, and $\mathtt{T}^X_{r_v} : \widetilde{\mathcal{X}} \rightarrow \widetilde{\mathcal{X}}$. And the same on the raw pose and motion space is represented as $\mathtt{T}^y_{r_z} : \mathcal{Y} \rightarrow \mathcal{Y}$, and $\mathtt{T}^Y_{r_v} : \widetilde{\mathcal{Y}} \rightarrow \widetilde{\mathcal{Y}}$. Note that, except $\mathtt{T}^z_{r_z}$ and $\mathtt{T}^v_{r_v}$, all other transformers are simple affine-like operations without involving neural network mappings.

a) Distilling lower-order non-local relations. For each target image $x_t \in \mathcal{X}$, we first construct a dataset of positive coupling represented as $\mathcal{N}^{\mathbf{x}}_{\mathbf{r_z}} = \{(x_t, x_t^+)$ where $x_t^+ = \mathtt{T}^x_{r_z}(x_t)\}$. Here, one is

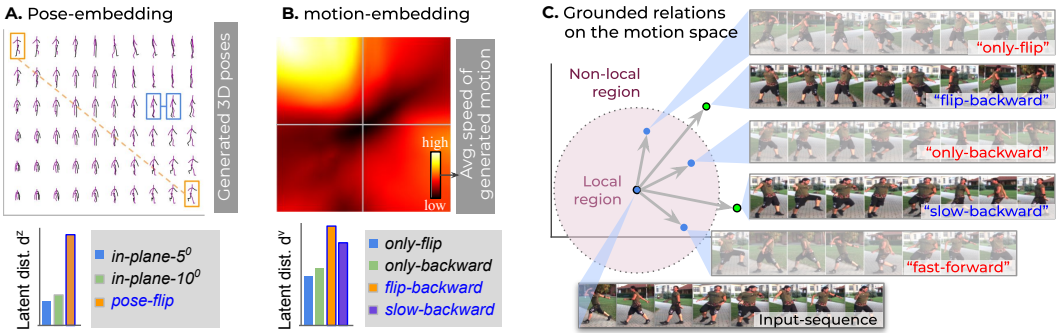

Figure 3: Embedding visualizations. **A.** Grid-interpolation of pose embedding (`pose-flip`: orange boxes; `in-plane-5°`: blue boxes) with bar-plot showing the avg. latent-distance. **B.** Grid-interpolation of speed of the generated motions with latent-distance bar-plot comparing the non-local motion relations. **C.** 2D PCA projected motion embedding points with glances of the anchor video (middle-point) and the corresponding transformed videos obtained by operating over various motion-related non-local relation candidates. Note that, only the relations in blue are selected because of their high non-localness indicated by the latent-distances d^z and d^v.

already aware of the corresponding pose relation, *i.e.* $y_t^+ = \mathtt{T}^y_{r_z}(y_t)$. And, the same relation in the latent pose space is expressed as $z_t^+ = \mathtt{T}^z_{r_z}(z_t)$.

Let us consider a non-local relation termed as "`pose-flip`" with rule-id $r_z = 1$. Here, $\mathcal{N}_1^x$ is obtained by pairing each target image with its flipped version obtained via a simple image-flip transformation operation, (*i.e.* $\mathtt{T}_1^x$). Note that, a similar pose space transformation would involve swapping the left side joints with that of the right (*i.e.* $\mathtt{T}_1^y$). And, the corresponding latent-transformer $\mathtt{T}_1^z$ is a multi-layer neural network which is trained to regress $z_t^+ = \mathtt{E_p} \circ \mathtt{T}_1^y(y_t)$ while feeding $z_t = \mathtt{E_p}(y_t)$ as the input. Here, y_t is sampled from the unpaired 3D pose data $\mathcal{Y}$. Next, we formalize the corresponding relational energy loss that uses the frozen latent-transformer $\mathtt{T}_1^z$ and is represented as,

$$\mathcal{L}_1^z = \sum_{(x_t, x_t^+) \in \mathcal{N}_1^x} \|\mathtt{T}_1^z \circ \mathtt{G}(x_t) - \mathtt{G}(x_t^+)\| \tag{2}$$

b) Distilling higher-order non-local relations. One can define higher order non-local relations on the hierarchical motion space to devise a similar relational energy loss.

Let us consider a non-local relation termed as "`flip-backward`" with rule-id $r_v = 1$. Here, $\mathcal{N}_1^{\mathbf{X}} = \{(X, X^+)$ where $X^+ = \mathtt{T}_1^X(X)\}$ is obtained by pairing each target sequence X with its `flip-backward` counterpart which is obtained via temporal reversal (backwards) of the image-flipped frames. Accordingly, the corresponding latent-transformer $\mathtt{T}_1^v$ is a multi-layer neural network trained to regress $v^+ = \mathtt{E_m} \circ \mathtt{E_p} \circ \mathtt{T}_1^{\mathtt{Y}}(Y)$ while feeding $v = \mathtt{E_m} \circ \mathtt{E_p}(Y)$ as the input, where Y is sampled from the unpaired 3D pose sequences $\widetilde{\mathcal{Y}}$. And, the corresponding relational energy loss is represented as,

$$\mathcal{L}_1^v = \sum_{(X, X^+) \in \mathcal{N}_1^X} \|\mathtt{T}_1^v \circ \mathtt{G}(X) - \mathtt{G}(X^+)\| \tag{3}$$

Finally, cross-modal alignment is achieved by simultaneously minimizing all the relational energies, *i.e.*

$$\mathcal{L} = \mathcal{L}_{CR} + \mathcal{L}_{NLR}; \text{ where } \mathcal{L}_{CR} = \mathcal{L}_{LCR} + \mathcal{L}_{HCR} \text{ and } \mathcal{L}_{NLR} = \sum_{r_z} \mathcal{L}_{r_z}^z + \sum_{r_v} \mathcal{L}_{r_v}^v \tag{4}$$

Next, we discuss certain insights which can help us to formalize effective non-local relations. We also answer what makes us choose `pose-flip` and `flip-backward` as suitable non-local relations.

3.3.3 What makes non-local relations more effective?

In the latent pose space, we conceptualize two different relation transformers viz. a) `pose-flip`, $\mathtt{T}_1^z$ and b) `in-plane-rotation`, $\mathtt{T}_2^z$. The corresponding relational energies are denoted as $\mathcal{L}_1^z$ and $\mathcal{L}_2^z$ respectively. Though minimizing $\mathcal{L}_1^z$ yields a substantial improvement over *Ours-LL* (just minimizing $\mathcal{L}_{CR}$ without $\mathcal{L}_{NLR}$), jointly minimizing both $\mathcal{L}_1^z$ and $\mathcal{L}_2^z$ does not yield much improvement on the final pose estimation performance.

Quantifying non-localness via latent-distance. We hypothesize that, "*relations coupling diverse samples (long-range interactions) lead to better cross-modal alignment*". To evaluate this hypothesis, we define a distance metric to apprehend the extent of non-localness of individual relations which is termed as *latent-distance*. Here, the *latent-distance* $\mathbf{d_1^z} = \text{mean}_{(z_t, z_t^+) \in \mathcal{N}_1^z} \|z_t - z_t^+\|$ measures the average distance between relation pairs $(z_t = \mathtt{E_p}(y_t), z_t^+ = \mathtt{E_p} \circ \mathtt{T}_1^y(y_t))$, for relation-id $r_z = 1$, where $y(y_t)$ is sampled from the unpaired 3D pose data $\mathcal{Y}$. In Fig. 3A, the bar-plot comparing d_1^z and d_2^z

(for relation-id $r_z = 2$) clearly highlight the gap in the degree of non-localness thus justifying the aforementioned observation regarding the performance gap. We perform a similar metric analysis to select the best set of relations defined on the motion space $\mathcal{V}$ (see Fig. 3B and 3C).

Finalizing the non-local relations. At the end, we take up a single pose relation, *i.e.* "`pose-flip`" and two diverse motion relations, *i.e.* "`flip-backward`" and "`slow-backward`" to realize our final cross-modal alignment framework (refer Fig. 3). With $\mathcal{L}_1^z$ being the relational energy specific to the pose space relation `pose-flip`, we denote $\mathcal{L}_1^v$ and $\mathcal{L}_2^v$ as the relational energies specific to the motion relations "`flip-backward`" and "`slow-backward`" respectively. The relational pairs for `slow-backward` are constructed by sampling the original sequence at 15 FPS, to enable a temporally slow 30 FPS slow backward video taken from the middle region (see Fig. 3C.

Overview of the optimization steps. Algorithm 1 summarizes the training as a step-by-step learning process. Here, each step uses frozen models obtained from the previous step. After initializing $\mathtt{G}$ from $\mathtt{G}^{\mathtt{s}}$ only a minimal set of network parameters (*Res-3* block of ResNet-50) are updated to learn target-specific mapping. This can be see as a regularization for the unsupervised adaptation training.

4 Experiments

We demonstrate effectiveness of the proposed framework by evaluating it on a variety of cross-dataset adaptation settings.

// A. Pre-learning steps (training on source data)
1) Train $\mathtt{E_p}$ and $\mathtt{D_p}$ on unpaired 3D poses, $y \in \mathcal{Y}$
2) Train $\mathtt{E_m}$ and $\mathtt{D_m}$ on unpaired motion, $Y \in \widetilde{\mathcal{Y}}$
3) Train $\mathtt{G}^{\mathtt{s}}$ on labeled source dataset, $(x^s, y^s) \in \mathcal{D}^s$
4) Train $\mathtt{T}_1^z$ for "`pose-flip`" on $\mathcal{Z}$ using $y \in \mathcal{Y}$
5) Train $\mathtt{T}_1^v$ for "`flip-backward`" on $\mathcal{V}$ using $Y \in \widetilde{\mathcal{Y}}$
6) Train $\mathtt{T}_2^v$ for "`slow-backward`" on $\mathcal{V}$ using $Y \in \widetilde{\mathcal{Y}}$

// B. Unsupervised target adaptation
7) After initializing $\mathtt{G}$ from $\mathtt{G}^{\mathtt{s}}$, train $\mathtt{G}$ by minimizing the local relational energy, $\mathcal{L}_{CR}$ alongside minimizing the non-local relations $\mathcal{L}_1^z$, $\mathcal{L}_1^v$ and $\mathcal{L}_2^v$.

Algorithm 1: Overview of the optimization steps. Note that, $\mathcal{L}_1^z$, $\mathcal{L}_1^v$ and $\mathcal{L}_2^v$ rely on frozen $\mathtt{T}_1^z$, $\mathtt{T}_1^v$ and $\mathtt{T}_2^v$ to distill the embedded relational *dark-knowledge.*

Implementation details. The backbone of $\mathtt{G}^{\mathtt{s}}$ constitutes of an ImageNet initialized ResNet-50 [23] (till *Res-4F*) followed by a series of fully-connected (FC) layers to obtain the latent pose representation, $z \in \mathbb{R}^{32}$. The pose auto-encoder, $\{\mathtt{E_p}, \mathtt{D_p}\}$ are FC networks operating on the 3D pose y (17 joints). The motion auto-encoder, $\{\mathtt{E_m}, \mathtt{D_m}\}$ is composed of bidirectional LSTMs [22] with 128 hidden units operating on a fixed sequence length of 30 (30 FPS) where the intermediate motion-embedding, $v \in \mathbb{R}^{128}$. The relation transformers constitutes of simple FC layers. We associate separate Adam optimizer [32] to each relational energy term which are optimized in alternate training iterations. The adaptation is performed on an Nvidia V100 GPU with each batch containing 8 videos each of frame-length 30 (see Suppl. for more details).

Datasets. We use the CMU-MoCap [1] dataset as the sample set for unpaired 3D poses $\mathcal{Y}$ and unpaired pose sequences $\widetilde{\mathcal{Y}}$. We use the synthetic SURREAL (**S**) dataset [74] as one of the source datasets. The sample set for the unpaired videos $\widetilde{\mathcal{X}}$ constitutes of single-person action videos (dance forms, sports, etc.) collected from the Sports-1M dataset [31]. The raw video frames are forwarded through a person-detector [59] to obtain the person-focused image sequences. We name it as *web-dataset* (**W**). For a fair evaluation, we use the standard, in-studio Human3.6M (**H**) dataset [26] as both source or target domain, in different problem settings. MPI-INF-3DHP (**M**) [45] is used as an unlabeled target domain to evaluate cross-studio adaptation. Further, 3DPW [75], and LSP [28] datasets are used to evaluate our cross-dataset generalizability, without involving these during training.

4.1 Adaptation settings

We introduce the following adaptation settings with various source and target dataset selections.

a) ***Ours(S→H)*** We use labeled synthetic SURREAL (S) as the source while unlabeled Human3.6M (H) acts as the target without involving the web-dataset (W). The resultant model is tuned to work well only for the specific in-studio environment, thus may fail to generalize for in-the-wild data.

b) ***Ours(H→M)*** Here, labeled Human3.6M is used as the only source domain while MPI-INF-3DHP (M) samples are used as the unlabeled target. This evaluates our cross-studio adaptation performance.

c) ***Ours(S→W)*** Here, labeled SURREAL (S) is used as the only source domain while the unlabeled target videos are extracted from the web-dataset (W). The resultant model is evaluated for cross-dataset generalization on 3DHP and 3DPW (see Table 3 and Table 4). Note that, web-dataset is the most challenging in-the-wild dataset with substantial diversity in pose, apparel, and backgrounds.

Table 2: Comparison of 3D pose estimation results on Human3.6M dataset. *Full (3D) supervision* denotes using GT 3D supervision for training. *Semi-Sup.* denotes supervised training only on subject S1. *Unsup.* refers to training on unpaired Human3.6M samples (unlabeled). We achieve SOTA results on both semi-supervised and unsup. training methods. Our unsupervised training outperforms the SOTA even in presence of huge domain gap between the SURREAL and H3.6M datasets. (MV) denotes usage of multi-view supervision.

Training	Methods	PA-MPJPE ↓	MPJPE ↓
Full (3D) Sup.	Chen *et al.* [10]	82.7	-
	Martinez *et al.* [44]	47.7	-
	Li *et al.* [37]	38.0	-
	Xu *et al.* [79]	36.2	45.6
	Chen *et al.* [14]	32.7	47.3
Semi-sup. (sup. on S1)	Mitra *et al.* [48]	90.8	120.9
	Li *et al.* [38]	66.5	88.8
	Rhodin *et al.* [60]	65.1	-
	Kocabas *et al.* [33]	60.2	-
	Iqbal *et al.* [27]$^{(MV)}$	51.4	62.8
	Ours(S→H, Semi)	**48.7**	**57.9**
Unsup.	Kundu et al. [36]	99.2	-
	Ours(S→H)	**86.8**	99.5

Table 3: Quantitative results for 3D pose estimation on 3DHP under three different supervision settings. *Full (3D) supervision* denotes using GT 3D supervision on 3DHP during training. *Unsupervised Adaptation* refers to training on labeled Human3.6M and adapting to unlabeled 3DHP. *Direct Transfer* denotes training on labeled source dataset, adapting to unlabeled web-dataset and evaluating on unseen 3DHP dataset. * denotes the implementation taken from [84]. (MV) denotes using multiview data.

Training	Methods	PCK ↑	AUC ↑	PA-MPJPE ↓
Full (3D) Sup.	Rogez *et al.* [62]	59.6	27.6	158.4
	Mehta *et al.* [46]	76.6	40.4	124.7
	Kocabas *et al.* [33]	77.5	-	-
	Iqbal *et al.* [27]	83.0	-	-
Unsup. Adapt.	Iqbal *et al.* [27]$^{(MV)}$	76.5	-	122.4
	Wandt *et al.* [76]*	81.6	47	95.4
	Kundu *et al.* [36]	83.2	**58.7**	97.6
	Zhao *et al.* [87]*	86.0	46.7	96.8
	Ours(H→M)	**89.2**	58.1	**80.4**
Direct Transfer	Chen *et al.* [11]	64.3	31.6	-
	Li *et al.* [37]	81.2	46.1	-
	Kundu *et al.* [36]	82.1	56.3	103.8
	Ours(S→W)	78.8	43.1	116.1
	Ours(SH→W)	**84.9**	**60.1**	**95.2**

d) ***Ours(SH→W)*** Here, the source domain is a combination of labeled SURREAL (S) and Human3.6M (H) datasets while the unlabeled web-dataset (W) acts as the target. This setting enjoys the advantage of strong source supervision on natural but in-studio Human3.6M alongside SURREAL.

4.2 Evaluation on benchmark datasets

In this section, we evaluate the proposed approach on the standard benchmark datasets. MPJPE and PA-MPJPE [26] denote standard mean per joint position error metric computed before and after Procrustes Alignment [21]. Following prior arts [45], we report PCK and AUC, i.e. percentage of correct keypoints and area under the curve respectively for the 3DHP dataset.

a) Evaluation on Human3.6M. Table 2 lists a comparison of *Ours(S→H)* against the prior arts. Under unsupervised training setup, *Ours(S→H)* outperforms the prior state-of-the-art (SOTA) by a significant margin. We attribute this improvement to our attempt to utilize domain randomization via SURREAL and the proposed cross-dataset alignment procedure. Under semi-supervised training, the target adaptation is supervised on a small subset of labeled samples (*i.e.* subject S1 in Human3.6M) alongside unsupervised alignment on the rest. The resultant model, *Ours(S→H, Semi)* also outperforms the prior semi-supervised works even in the absence of additional multi-view supervision as used in some of the prior-arts.

Table 4: Quantitative results comparing our approach against prior-arts for 3D pose estimation on 3DPW. *Full (3D) supervision* denotes using ground-truth 3D supervision on 3DPW for training. *Direct Transfer* denotes training on a variety of labeled source dataset and directly evaluating the resultant model on unseen 3DPW test set. We achieve state-of-the-art result upon using both Human3.6M (H) and SURREAL (S) as the source and the web-data as the target. $^{+}$ denotes number taken from [30]. * denotes using parametric mesh models.

Training	Methods	PA-MPJPE ↓
Full (3D) Supervision	Arnab *et al.* [3]*	77.2
	Sun *et al.* [69]*	69.5
Direct Transfer	Martinez *et al.* [44]$^{+}$	157.0
	Dabral *et al.* [15]$^{+}$	92.3
	Kanazawa *et al.* [30]*	80.1
	Doersch *et al.* [16]*	82.4
	Kanazawa *et al.* [29]*$^{+}$	76.7
	Ours(S→W)	79.7
	Ours(SH→W)	**72.9**

b) Evaluation on 3DHP. Table 3 shows a detailed quantitative comparison of three of our adaptation variants; *Ours(H→M)*, *Ours(S→W)*, and *Ours(SH→W)*, under two training setups. *Ours(SH→W)* achieves state-of-the-art unseen transfer performance validating superior generalizability of the proposed framework. *Ours(SH→W)* uses in-studio but real Human3.6M as an additional labeled source data alongside SURREAL, thus reducing the domain gap when compared against *Ours(S→W)*.

c) Evaluation on 3DPW. Table 4 reports a detailed comparison of our variants against the prior-arts on the in-the-wild 3DPW dataset. All the methods under *direct transfer* do not use 3DPW samples even for any supervised or unsupervised training. A lower PA-MPJPE for *Ours(SH→W)* clearly

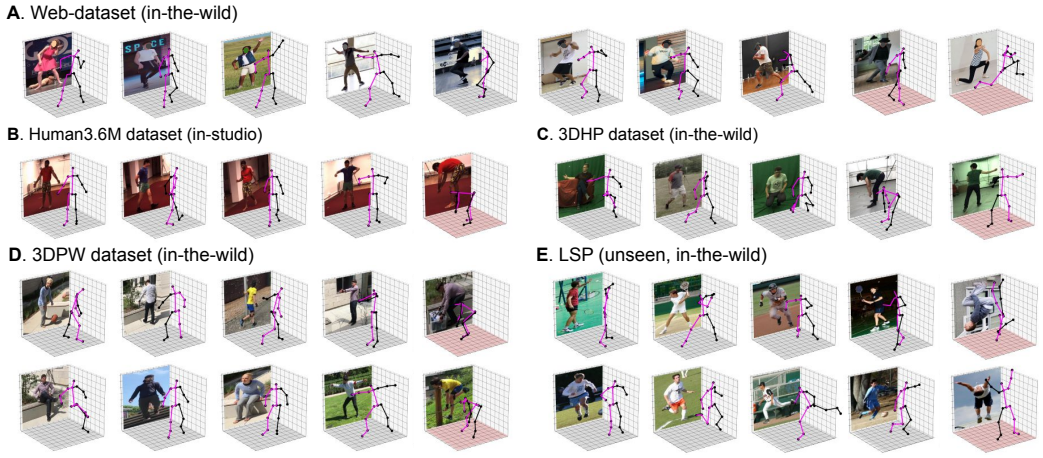

Figure 4: Predictions obtained via *Ours(SH→W)* generalize well to both in-studio and in-the-wild datasets. The failure cases (rare poses or poses with inter-limb occlusion) are highlighted by results on red bases.

highlight our superior cross-dataset generalizability against the prior approaches which also utilize additional information such as the SMPL mesh model [41].

d) Ablation study on Human3.6M. Table 5 shows an ablative analysis highlighting the effectiveness of individual relational energies towards realizing a better unsupervised alignment. As compared to the source-only baseline (first row of Table 5), distillation of contrastive relations in $\mathcal{L}_{\text{CR}}$ yields the first stage of improvement. Minimizing the non-local pose space relational energy $\mathcal{L}_1^z$ results in further improvement by effectively distilling an instance-grounded *dark-knowledge*. Additionally, distilling higher order non-local motion relations (*i.e.* $\mathcal{L}_1^v$ and $\mathcal{L}_2^v$) yields a substantial improvement by effectively attenuating the instance-level misalignment.

Table 5: Ablation study on Human3.6M dataset (*i.e.* *Ours(S→H)*). Starting from the baseline (inference through the SURREAL trained model), usage of different relational energies results in a substantial improvement in the unsupervised adaptation performance.

Source-only training	Ablation on target				MPJPE ↓
	$\mathcal{L}_{CR}$	$\mathcal{L}_1^z$	$\mathcal{L}_1^v$	$\mathcal{L}_2^v$	
✓	-	-	-	-	209.6
✓	✓	-	-	-	172.1
✓	✓	✓	-	-	141.2
✓	✓	✓	✓	-	95.6
✓	✓	✓	✓	✓	86.8

Qualitative evaluation and limitations. Fig. 4 illustrates the generaliability of *Ours(SH→W)* proposed approach under diverse pose, apparel, and background variations. Results with red base show some of the failure cases. The model is restricted from predicting implausible 3D pose outcomes as these are inferred through the prior-enforcing generative pose decoder. However, the model fails under certain drastic scenarios such as high background clutter, multi-level body-part occlusion, and rare athletic poses. Other limitations include body-truncation scenarios *i.e.*, scenarios where certain body-parts are either occluded by external objects or are outside the image frame. In such cases there are multiple plausible 3D pose outcomes, thus asking for a future exploration of probabilistic pose estimation modeling. Taking a different direction, one can explore innovative ways to utilize inputs from auxiliary modalities (such as depth, foreground segmentation, etc.) as and when available from the deployed target environment (see Suppl. for more details).

Negative societal impacts. While we do not foresee our framework causing any direct negative societal impact, it may be leveraged to create malicious applications utilizing human tracking and action recognition. Estimating a human pose does not require any identity information. However, methods such as gait recognition can be indirectly used to identify personal attributes thereby raising privacy concerns. We urge the readers to make use of the work detailed responsibly, limiting the usage to legal use-cases.

5 Conclusion

We presented a cross-modal alignment technique to align the learned representations from two diverse modalities. Our unsupervised technique allows adaptation to the wildest unlabeled samples gathered from web while initializing the base model on substantially diverse but unrealistic SURREAL. We analyzed the importance of higher order, non-local relations in expressing the rich instance-grounded dark knowledge as required to attenuate the instance-level misalignment. In future, we plan to extend our framework for multi-modal alignment in presence of additional input modalities.

References

[1] CMU graphics lab motion capture database. available: http://mocap.cs.cmu.edu/. 7

[2] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *Transactions on Pattern Analysis and Machine Intelligence*, 2006. 3

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019. 8

[4] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 3

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

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

[7] Federica Bogo, A. Kanazawa, Christoph Lassner, P. Gehler, J. Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 3

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

[9] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *ICCV*, 2019. 3

[10] Ching-Hang Chen and Deva Ramanan. 3D human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2017. 8

[11] 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, 8

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 4

[13] 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. 3

[14] Zerui Chen, Yan Huang, Hongyuan Yu, Bin Xue, Ke Han, Yiru Guo, and Liang Wang. Towards part-aware monocular 3d human pose estimation: An architecture search approach. In *ECCV*, 2020. 8

[15] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3D human pose from structure and motion. In *ECCV*, 2018. 8

[16] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In *NeurIPS*, 2019. 3, 4, 8

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 4

[18] Dedre Gentner and Kenneth J. Kurtz. Relational categories. In *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin.* 2005. 2

[19] Michael Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998. 1

[20] Micah B. Goldwater, Hilary J. Don, Moritz J. F. Krusche, and Evan J. Livesey. Relational discovery in category learning. *Journal of Experimental Psychology: General*, 2018. 2

[21] John C Gower. Generalized procrustes analysis. *Psychometrika*, 1975. 8

[22] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 2005. 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 4

[25] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d human pose estimation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018. 3

[26] 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, 7, 8

[27] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, 2020. 3, 8

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

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

[30] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 8

[31] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 7

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[33] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3D human pose using multi-view geometry. In *CVPR*, 2019. 1, 8

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

[35] Jogendra Nath Kundu, Ambareesh Revanur, Govind V Waghmare, Rahul M Venkatesh, and R Venkatesh Babu. Unsupervised cross-modal alignment for multi-person 3d pose estimation. 2020. 3

[36] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *CVPR*, 2020. 8

[37] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *CVPR*, 2020. 8

[38] Z. Li, X. Wang, F. Wang, and P. Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *ICCV*, 2019. 8

[39] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *CVPR*, 2019. 2

[40] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 4

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics*, 2015. 9

[42] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019. 2, 5

[43] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 1, 3

[44] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 3, 8

[45] 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, 7, 8

[46] 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. 8

[47] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 2, 4

[48] Rahul Mitra, Nitesh B. Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *CVPR*, 2020. 8

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

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

[51] Kent L Norman. *Cyberpsychology: An introduction to human-computer interaction*. Cambridge university press, 2017. 1

[52] 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

[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 3, 5

[54] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 2

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

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

[57] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 3

[58] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *CVPR*, 2019. 2

[59] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 7

[60] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *ECCV*, 2018. 3, 8

[61] 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

[62] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 8

[63] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2, 4

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

[65] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 4

[66] 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

[67] Suraj Srinivas and Francois Fleuret. Knowledge transfer with jacobian matching. In *ICML*, 2018. 2

[68] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018. 3

[69] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019. 8

[70] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2019. 2

[71] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *CVPR*, 2019. 2

[72] 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

[73] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 4

[74] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 1, 7

[75] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2, 7

[76] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In *CVPR*, 2019. 1, 3, 8

[77] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, 2011. 4

[78] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. 2020. 3

[79] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *CVPR*, 2020. 8

[80] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 1

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

[82] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 2

[83] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2

[84] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3d human pose estimation. In *NeurIPS*, 2020. 8

[85] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3d human pose estimation. In *NeurIPS*, 2020. 3

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

[87] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, 2019. 3, 8

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Refer to Sec. 4.2
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Refer to Sec. 4.2
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Refer supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See implementation details in Sec. 4 and supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] The metrics are reported in line with standard 3D human pose estimation protocols that do not involve any random seeds as multiple runs would be too computationally expensive. The feature extractor is initialized to standard ImageNet trained weights.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] in Sec. 4 and supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Sec. 4
   (b) Did you mention the license of the assets? [Yes] Refer to Supplementary material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Refer Supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use standard datasets as used in prior works. The papers introducing the datasets include additional details.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We utilize standard datasets collected responsibly, and have made sure to blur the images obtained from the web-dataset (*i.e.* selected samples from the Sports-1M dataset).
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]