Indian Institute of Science
Bangalore, India
भारतीय विज्ञान संस्थान
बंगलौर, भारत

NeurIPS 2021

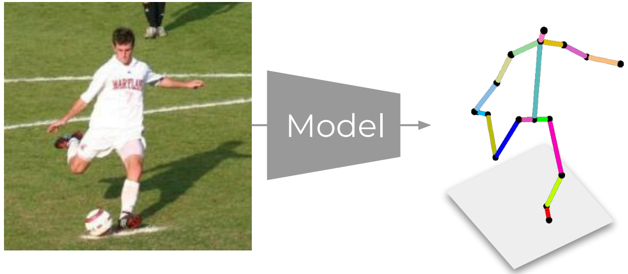
Non-local Latent Relation Distillation for Self-Adaptive 3D Human Pose Estimation

Jogendra Nath Kundu[1] Siddharth Seth[1] Anirudh Jamkhandi[1]
Pradyumna YM[1] Varun Jampani[2] R. Venkatesh Babu[1] Anirban Chakraborty[1]

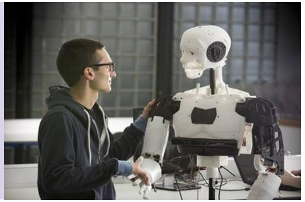
[1]Indian Institute of Science [2]Google Research

Goal task: 3D human pose estimation

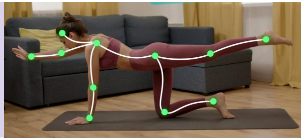
- Inferring 3D human pose from monocular RGB images.
- Key step to several human centric applications such as human-computer interaction, sports analytics, driver assistance, etc.



Human-robot interaction

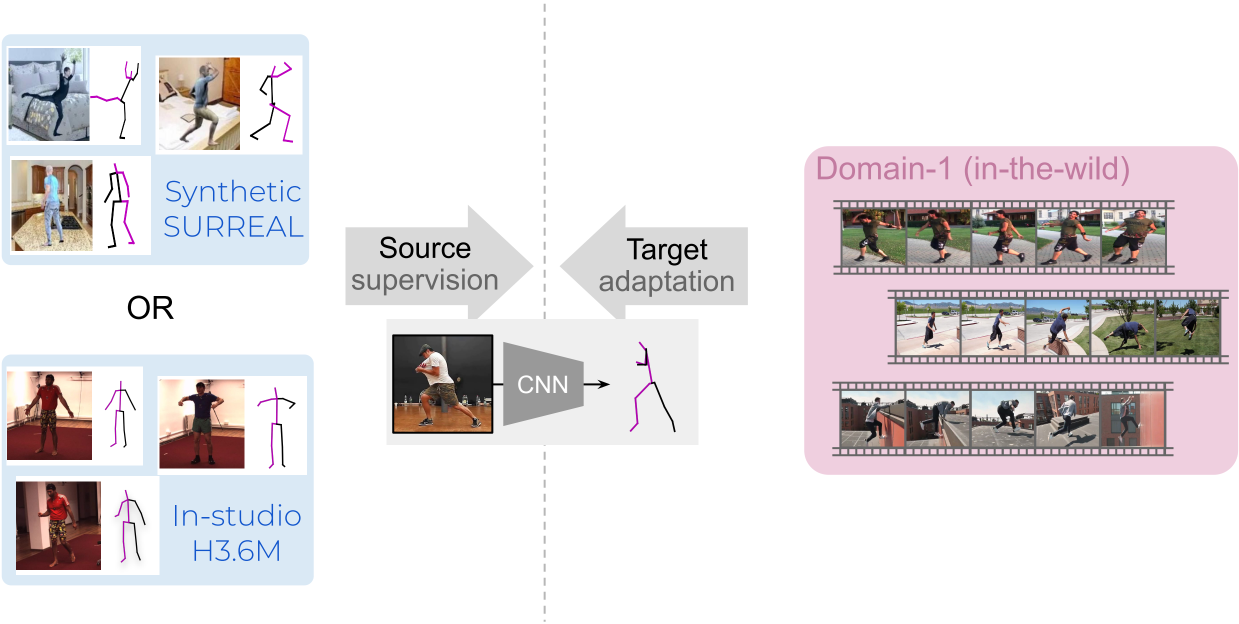


AI fitness tutor



AR/VR application

Domain adaptation: improving deployability of available solution

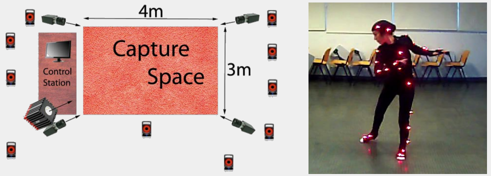


Target Label requirement
(reducing sup. levels)

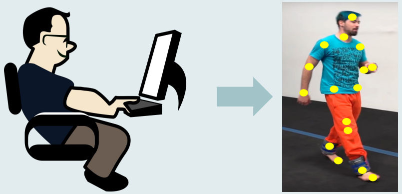
1. 3D pose GT
2. 2D pose GT
3. Unsupervised

One must minimize the target label requirements for convenient deployment.

Domain adaptation: improving deployability of available solution



Available 3D pose capturing systems are
difficult to configure in outdoors



Gathering 2D pose annotation
involves manual intervention

Target Label requirement
(reducing sup. levels)

1. 3D pose GT ✗
2. 2D pose GT ✗
3. Self-adaptive ✓

Self-adaptive: digress from any form of paired supervision or auxiliary cues.

We seek answers to the following.

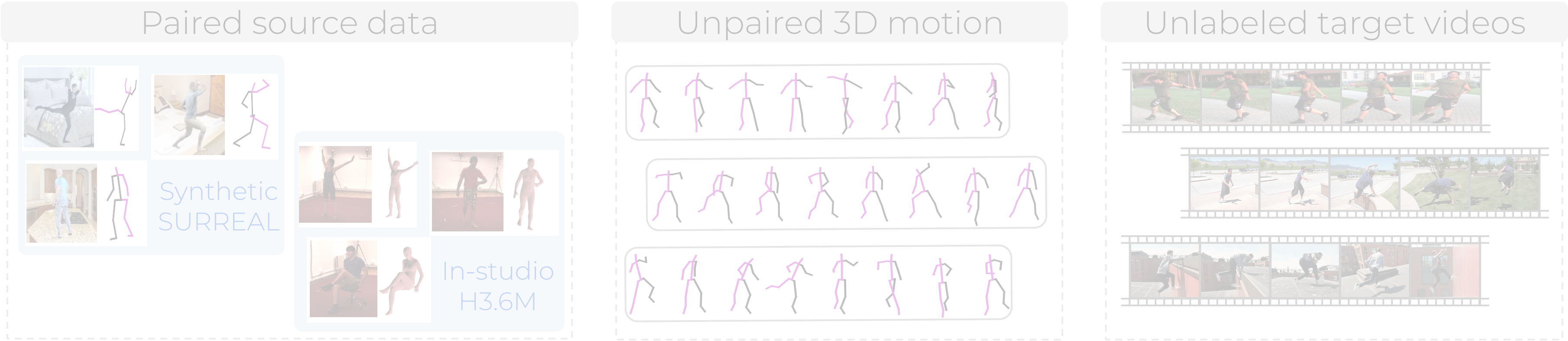
- Can we completely move away from paired supervision or auxiliary cues (multi-view or depth)?
- Can we develop a self-adaptive framework to avoid the curse of dataset-bias thereby aiming to attain superior cross-domain generalization?

We cast 3D pose learning as a self-supervised adaptation problem.

- We aim to transfer the task knowledge from a labeled source domain to a completely unlabeled target.

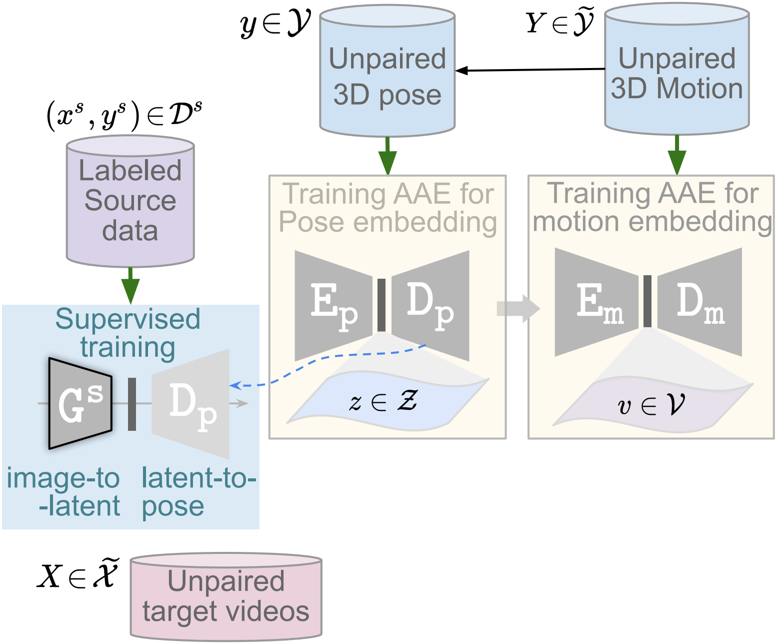
In the proposed setting we consider access to the following:

1. A labeled source dataset: either synthetic (SURREAL) or in-studio (Human3.6M) environment.
2. A dataset of unpaired 3D pose sequences.
3. A dataset of unlabeled video sequences from the target domain.



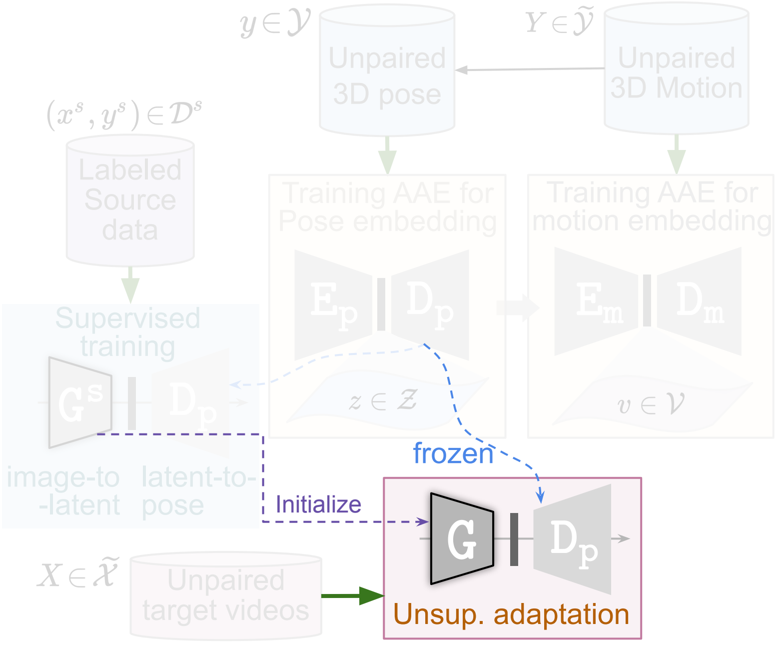
Overview: notations and modules

- Notations of the 3 datasets.
- We introduce 2 latent embeddings
  (learned via adv. auto-encoder)
  a) Pose embedding
  b) Motion embedding
- *Image-to-pose* inference is carried out via:
  a) *Image-to-latent*
  b) *Latent-to-pose*
- Supervised pre-learning of G^s
  (uses D_p as the latent-to-pose mapping)



Overview: notations and modules

- Notations of the 3 datasets.
- We introduce 2 latent embeddings
  (learned via adv. auto-encoder)
  a) Pose embedding
  b) Motion embedding
- *Image-to-pose* inference is carried out via:
  a) *Image-to-latent*
  b) *Latent-to-pose*
- Supervised pre-learning of $\mathtt{G}^{s}$
  (uses D_p as the latent-to-pose mapping)



Objective: Train image-to-latent **G** on unpaired target image sequences.

Approach: Distilling local neighborhood relations via contrastive learning

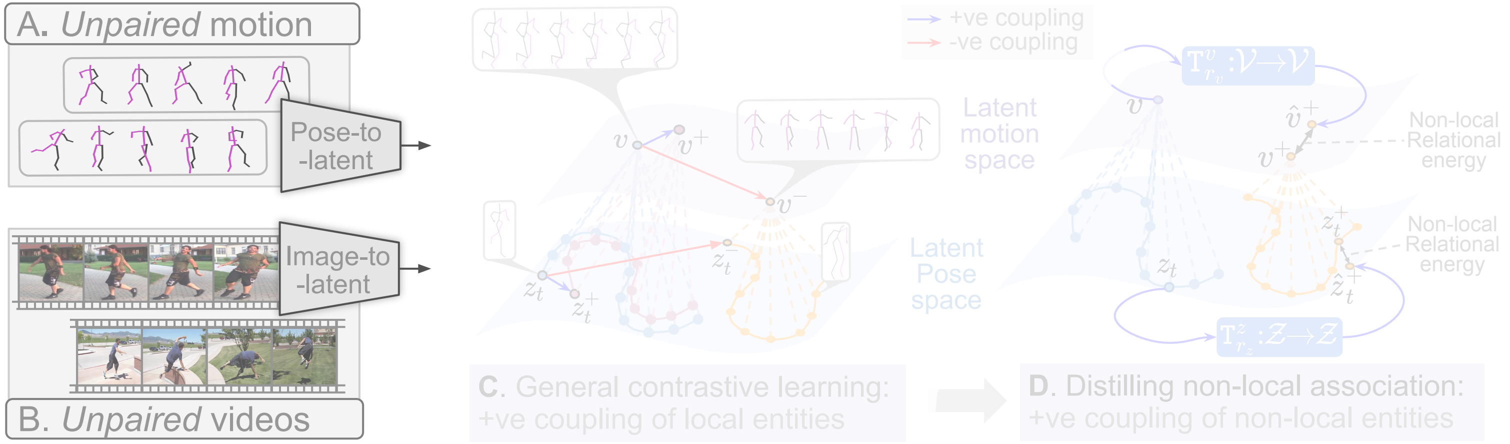
a) Lower-order contrastive (operating on pose-space)

b) Higher-order contrastive (operating on motion-space)

+ve coupling: pose-invariant image augmentations

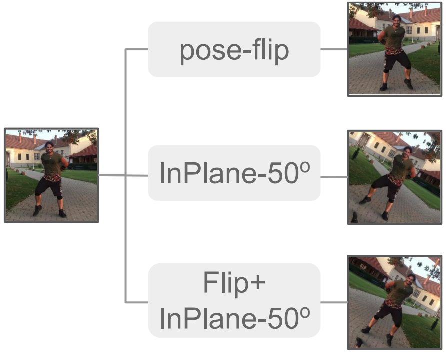
-ve coupling: random unrelated pose

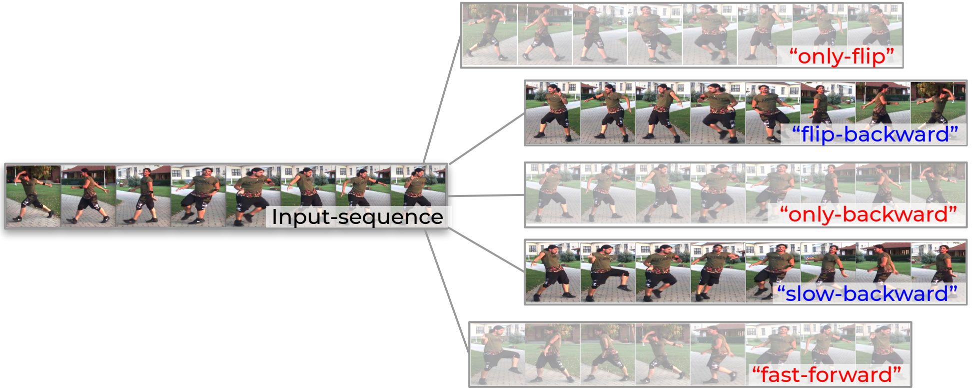
Why to use motion embedding when the goal task is to realize an image-to-pose mapping?



What are non-local relations?

- Non-local pose relations
- Non-local motion relations

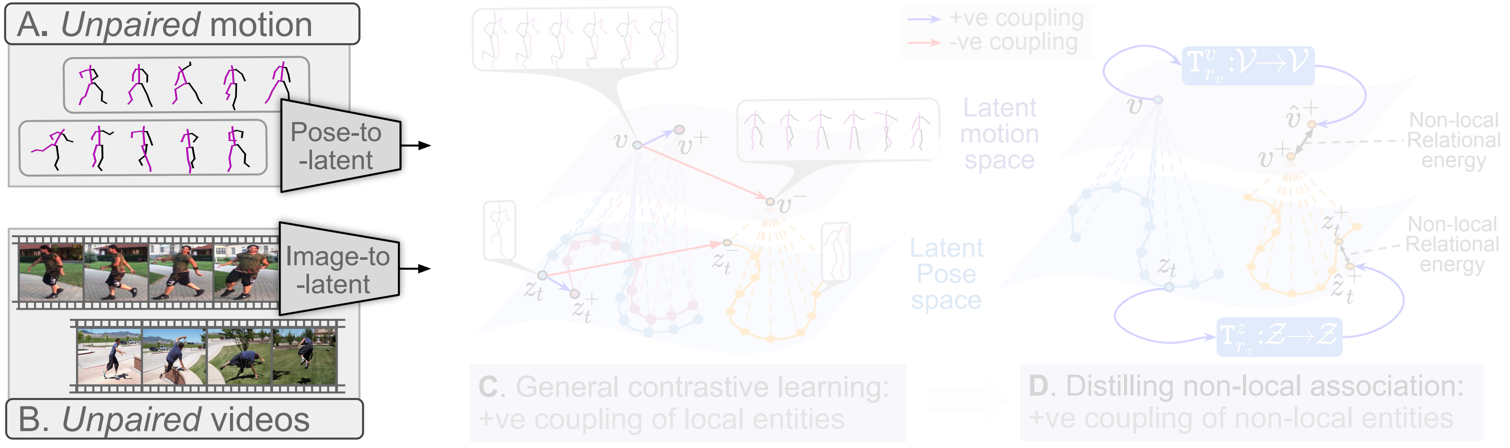




Approach: Distilling non-local relations via equivariance consistency

- Unlike contrastive relations non-local positive couplings characterize long-range latent pose/motion interactions.
- We propose to distill non-local relations via pre-learned relation transformer networks.

The equivariance consistency aims to preserve the equivariance of higher order spatio-temporal relations between the two modalities as a means to perform the cross-modal alignment.



What makes non-local relations more effective?

- Quantifying non-localness via latent-distance
- We show that relations coupling diverse samples (long-range interactions) lead to better cross-modal alignment

Results: adaptation from Synthetic to Real

Ours(S→H)

Syn. SURREAL → Real H3.6M

Evaluated on Human3.6M

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	-
	[illegible] (MV)	[illegible]	[illegible]
	Ours(S→H, Semi)	**48.2**	**57.6**
Unsup.	Kundu et al. [36]	99.2	-
	Ours(S→H)	**86.2**	97.8

against **unsupervised** prior-arts

[1] Kundu *et al.*, “Self-Supervised 3D Human Pose Estimation via Part Guided Novel Image Synthesis”, CVPR ‘20

Results: adaptation from Synthetic to Real

Ours(S→H, Semi)

Syn. SURREAL → Real H3.6M → Evaluated on Human3.6M

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	-
	[illegible]	[illegible]	[illegible]
	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)	**48.2**	**57.6**
Unsup.	Kundu et al. [36]	99.2	-
	Ours(S→H)	86.2	97.8

against **semi-supervised** prior-arts

[1] Iqbal *et al.*, "Weakly-Supervised 3D Human Pose Learning via Multi-view Images in the Wild", CVPR '20

Results: adaptation from Synthetic to Real

Ours(S→W)

Ours(SH→W)

against prior-arts on **unseen** 3DPW

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)	79.3
	Ours(SH→W)	**72.1**

Ablation experiments

Modules Involved:

- G - Image-to-latent model
- D_p - Frozen pose decoder
- E_m - Frozen motion encoder
- T_1^z - Flip+InPlane-50°
- T_1^v - Flip-backward+InPlane-50°
- T_2^v - slow-backward

Ablation study on Human3.6M

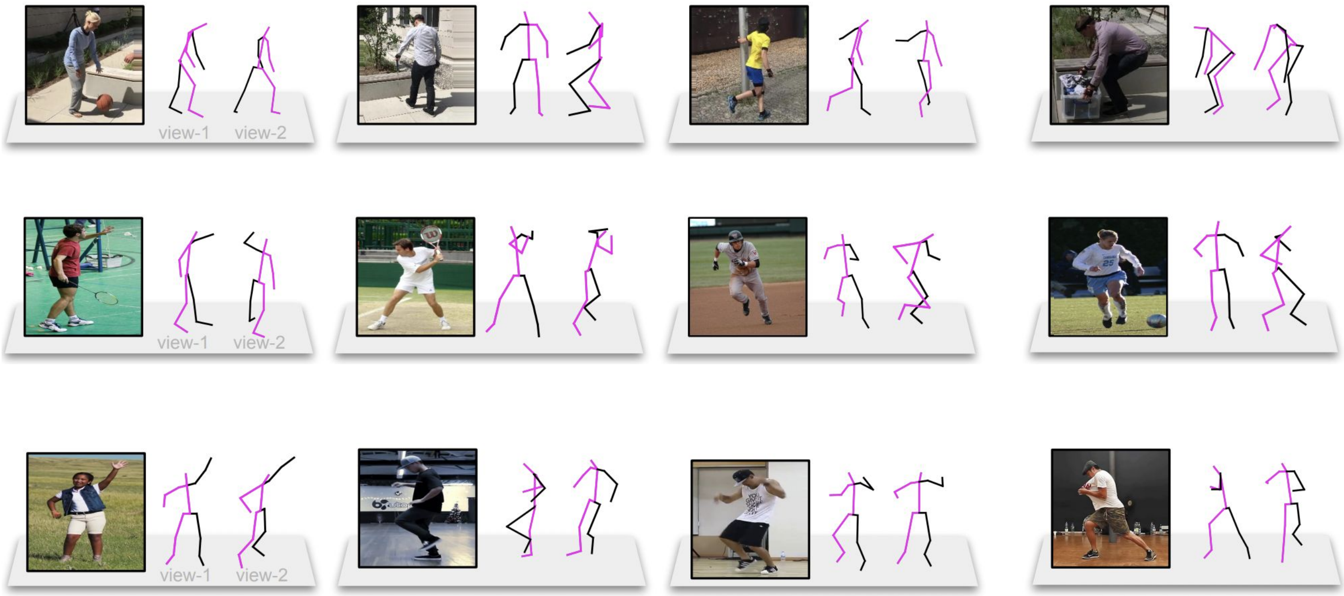
Ablation	Modules Involved	MPJPE ↓
Source-only	G, D_p	209.6
$+\mathcal{L}_{LCR}$	G, D_p	193.4
$+\mathcal{L}_{HCR}$	$+E_m$	172.1
$+\mathcal{L}_1^z$	$+T_1^z$	139.7
$+\mathcal{L}_1^v$	$+T_1^v$	91.8
$+\mathcal{L}_2^v$	$+T_2^v$	86.2

Qualitative Results: adaptation from Synthetic to Real

3DPW

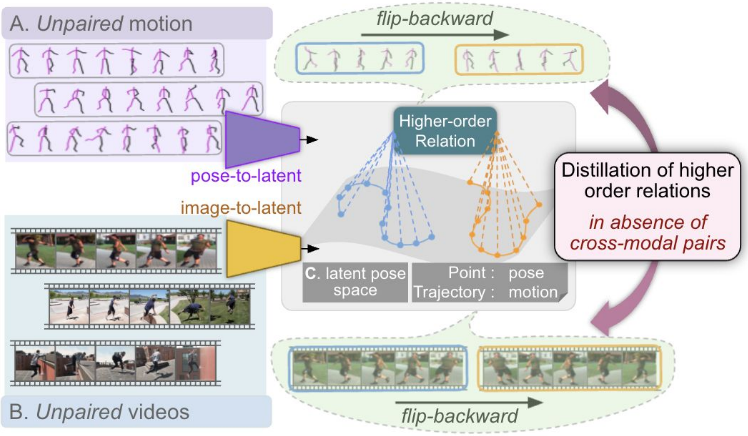
LSP

Web Dataset



Summary

- Our cross-modal alignment technique aligns the learned representations from two diverse modalities.
- Higher-order relations operating in motion space couple many entities → better cross modal alignment
- Non-local relations couple entities beyond structural neighborhood unlike in general contrastive learning.
- Latent distance objectively quantifies non-localness to select the most effective relation set.



Thank You!

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

Please check our project page for more details

https://sites.google.com/view/sa3dhp