Learning 3D Human Pose Estimation from Dozens of Datasets using a Geometry-Aware Autoencoder to Bridge Between Skeleton Formats

WACV WAIKOLOA HAWAII JAN 3-7 • 2023

István Sárándi, Alexander Hermans, Bastian Leibe

RWTH AACHEN UNIVERSITY

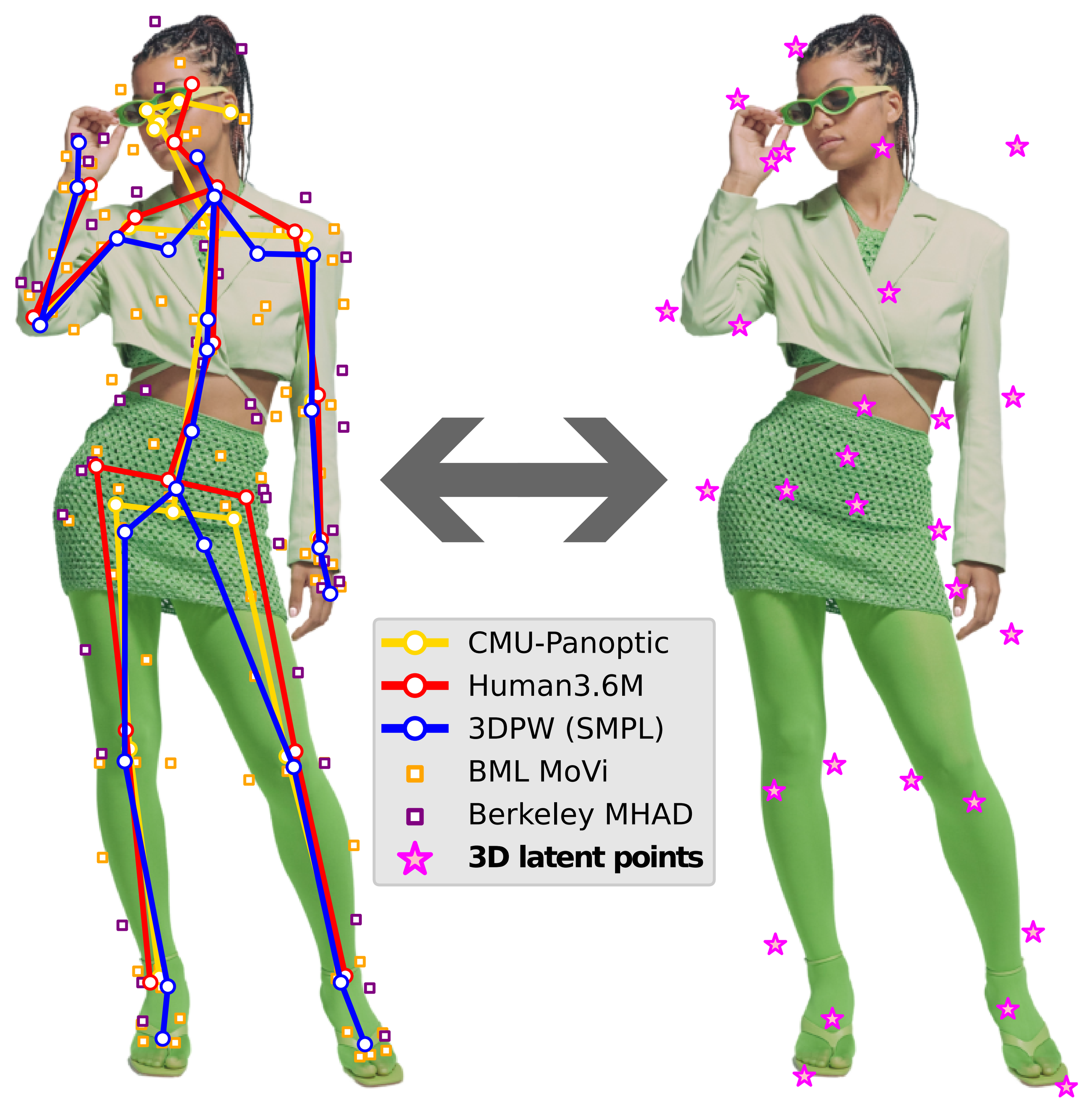
TL;DR *We discover how various 3D human* ***skeleton formats*** *are related, via a novel* ***affine-combining autoencoder****, enabling* ***extreme multi-dataset*** *3D pose training. We release strong 3D pose estimators for downstream research.*

vision.rwth-aachen.de/wacv23sarandi

The Problem

- Single RGB image → 3D human pose
- **Lack of data** is seen as a big problem
- This led many to focus on 2D-to-3D lifting and self-supervision
- Instead, we **push the fully-supervised regime** to the extreme
- Actually, **lots of datasets** exist now – let's use **28!** (see bottom)

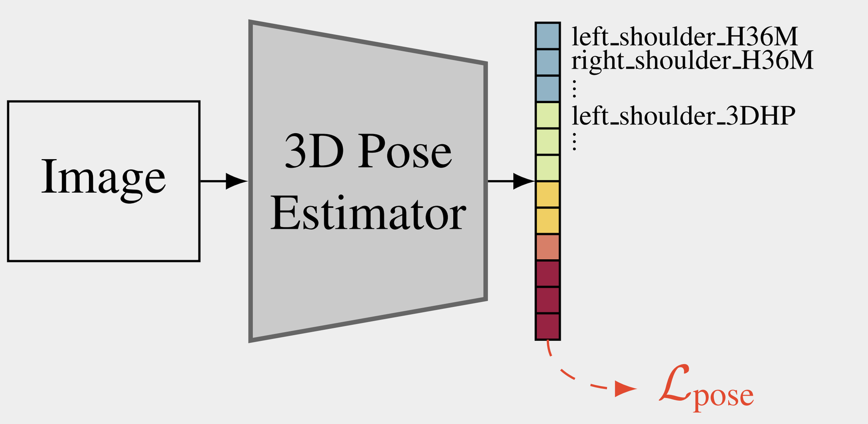
- But datasets use **different skeleton formats!** →
- Unclear how to **supervise one model** with them
- Idea: learn **how the skeleton definitions relate**
- Goal: effective information sharing across formats



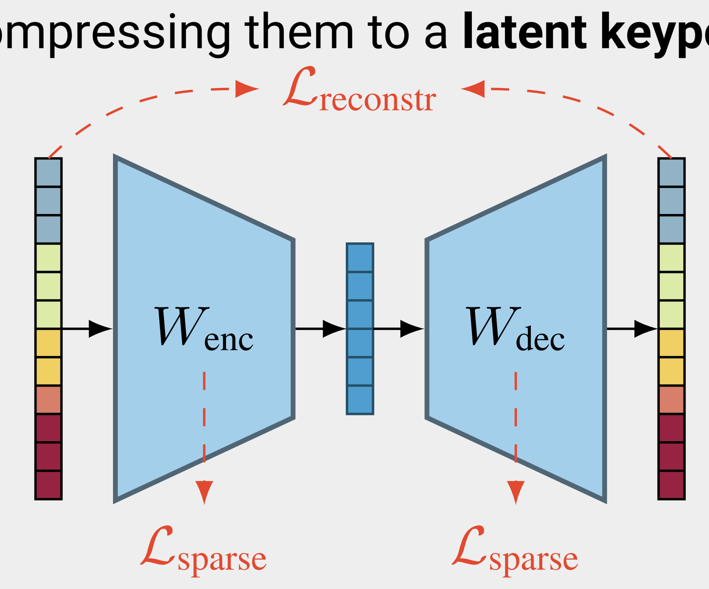
Approach

- Discover **latent 3D points** that best explain all formats
- Design a novel but simple **geometric autoencoder** for this
- Use pseudo-GT with **multiple formats per sample** to train autoencoder

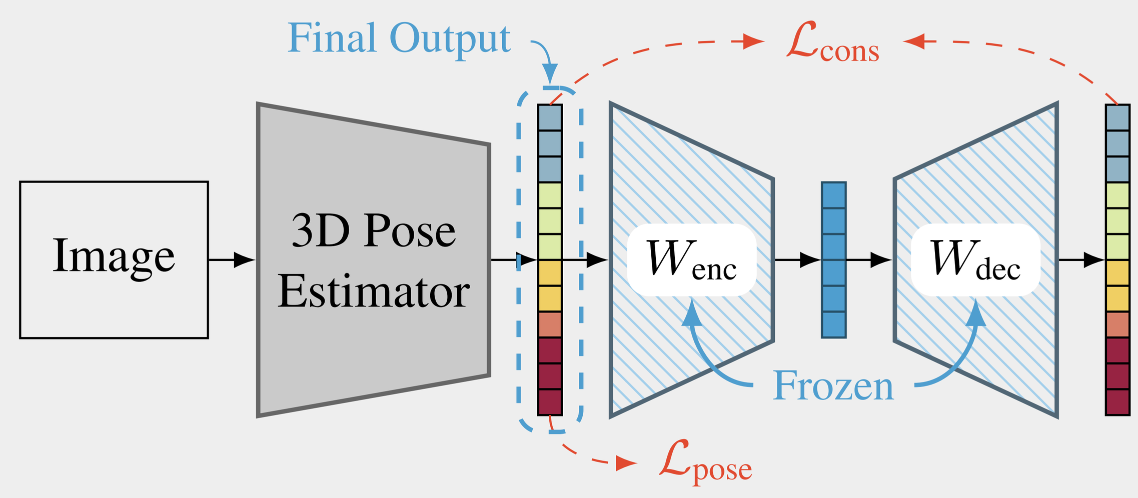
Step 1) Train a multi-dataset model with separate output heads per skeleton format, **to pseudo-annotate images with all formats**



Step 2) Train our **affine-combining autoencoder** to discover relations between formats by compressing them to a **latent keypoint set**



Step 3) Consistency-regularization: make the model predict poses near the latent space (alternatively: perform direct latent prediction)



Affine-Combining Autoencoder (ACAE)

- **Want: equivariance** to rotation, translation, scale, chirality
- **Linear, constrained, regularized autoencoder**
- Constraint 1) same weights for X, Y, Z coordinates
- Constraint 2) weights sum to one (affine combination)
- Regulariziation loss: L1 for sparsity
- Reconstruction loss: L1
- **Generally applicable** to compress large sets of keypoints

$$\underset{W_{\text{enc}}\in\mathbb{R}^{L\times J},\, W_{\text{dec}}\in\mathbb{R}^{J\times L}}{\text{minimize}} \quad \mathcal{L}_{\text{reconstr}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}}$$

$$\mathcal{L}_{\text{reconstr}} = \frac{1}{K}\sum_{k=1}^{K} \left\| P_k - W_{\text{dec}} W_{\text{enc}} P_k \right\|_1$$

$$\mathcal{L}_{\text{sparse}} = \|W_{\text{enc}}\|_1 + \|W_{\text{dec}}\|_1$$

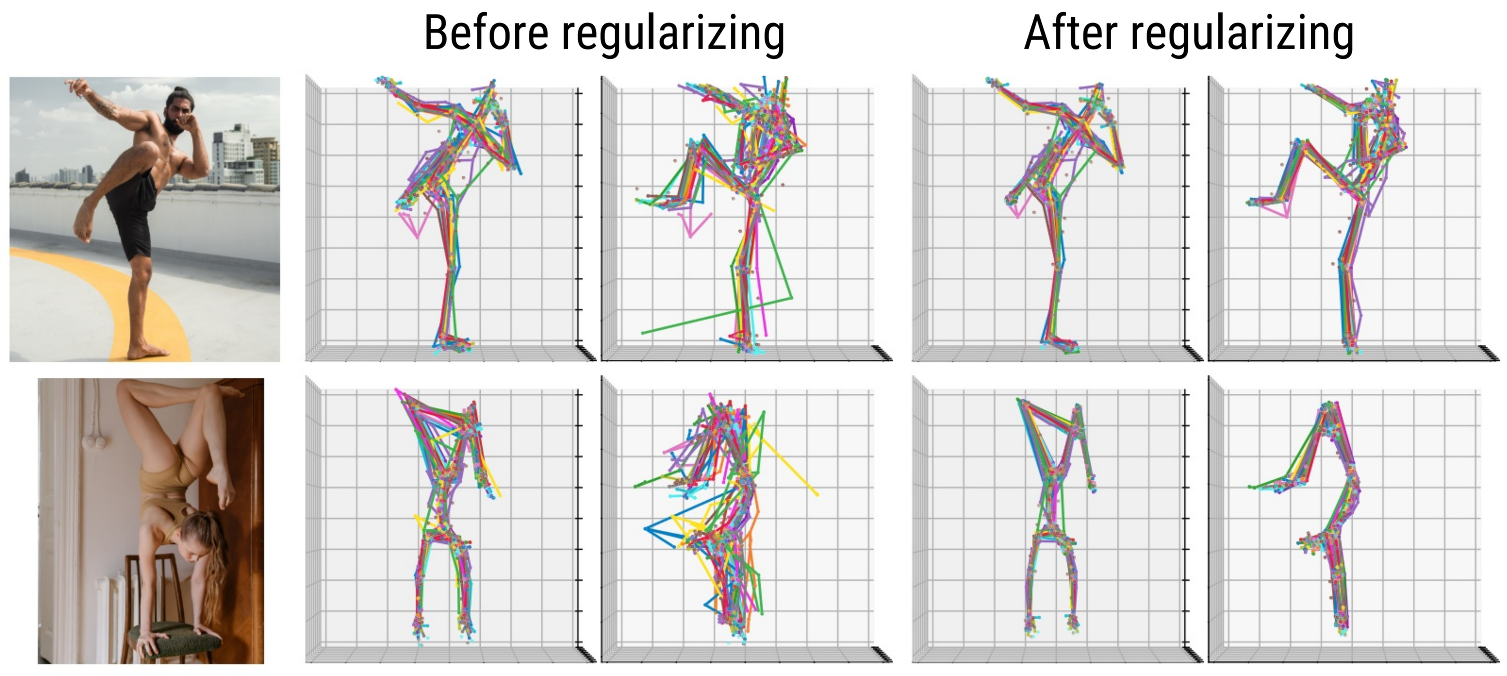
$$\text{s. t.} \quad W_{\text{enc}}\mathbf{1}_J = \mathbf{1}_L, \quad W_{\text{dec}}\mathbf{1}_L = \mathbf{1}_J,$$

Findings

- Base model is our **MeTRAbs** pose estimator
- **Data scale helps!** Steady improvement with 1→ 3 → 14 → 28 ds.
- Separate skeleton output heads give inconsistent depth predictions
- **ACAE consistency-regularization improves consistency**
- **Models become much stronger** than those from prior work
- Models available at `vision.rwth-aachen.de/wacv23sarandi`

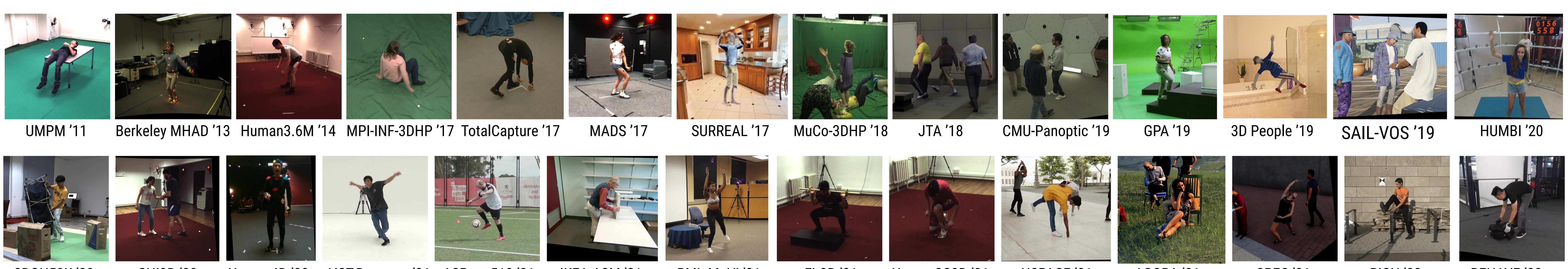
		MuPoTS-3D				3DPW				MPI-INF-3DHP				Human3.6M			
		MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑
EffNetV2-S	Merged joints	91.9	67.3	63.2	69.9	72.5	48.3	79.5	69.7	69.8	51.6	80.7	79.4	**44.6**	34.2	93.9	89.8
	Separate joints (F. 3a)	84.6	59.0	70.1	66.0	61.8	43.4	83.8	71.1	59.6	44.1	**86.6**	81.8	44.7	34.3	94.3	**90.1**
	Consistency regul. (F. 3c)	**81.8**	**57.8**	**72.5**	**72.9**	**61.5**	**43.0**	**84.0**	**71.9**	**59.2**	**43.6**	**86.6**	**82.7**	45.2	**33.3**	**94.4**	**90.1**
	Latent pred. (F. 4a)	83.0	58.9	71.4	71.2	62.0	43.6	**84.0**	71.7	60.2	44.7	86.1	80.2	46.5	34.4	93.9	89.5
	Hybrid (F. 4b)	82.7	58.5	71.6	72.1	61.8	43.3	**84.0**	71.8	60.4	44.8	85.9	80.9	46.1	34.2	94.1	89.4
EffNetV2-L	Separate joints	82.9	57.7	71.0	70.9	60.9	42.1	84.4	73.4	59.1	42.2	88.0	**85.3**	41.6	32.0	95.1	92.1
	Consistency regul.	**81.0**	**57.4**	**72.8**	**74.8**	**60.6**	**41.7**	**84.7**	**74.3**	**57.9**	**41.8**	**88.2**	84.7	**40.6**	**30.7**	**95.7**	**92.6**
	Hybrid	81.3	57.9	72.4	73.9	61.1	42.0	84.6	**74.3**	59.2	42.8	87.2	84.3	41.8	31.4	95.6	**92.6**

Before regularizing | After regularizing



	MuPoTS-3D	3DPW			MPI-INF-3DHP		Human3.6M
	PCK_{150}↑	MPJPE↓	PMPJPE↓	PCK_{50}↑	MPJPE↓	PCK_{150}↑	MPJPE↓
Sun *et al.* (2021)	–	80.1	56.8	36.5	–	–	–
Lin *et al.* (2021b)	–	74.7	45.6	–	–	–	51.2
Gong *et al.* (2021)	–	–	–	–	71.1	89.2	50.2
Cheng *et al.* (2022)	89.6	–	–	–	–	–	49.3
Ours with crop resolution 256x256 and 400k steps							
ResNet-50	92.2	65.5	47.2	49.0	64.2	93.3	45.8
EffNetV2-S	93.7	61.5	43.0	51.8	60.0	95.3	45.2
EffNetV2-L	94.1	60.6	41.7	52.1	59.2	95.8	40.6
Ours with crop resolution 384x384 and 800k steps							
EffNetV2-S	94.9	59.5	41.0	53.1	58.7	96.2	41.4
EffNetV2-S 5-crop TTA	95.2	58.9	39.9	53.6	57.5	96.7	40.1
EffNetV2-L	95.4	58.9	39.5	53.9	55.4	97.1	36.5
EffNetV2-L 5-crop TTA	95.7	57.0	38.1	55.4	53.6	97.6	35.5

The Used Datasets



UMPM '11, Berkeley MHAD '13, Human3.6M '14, MPI-INF-3DHP '17, TotalCapture '17, MADS '17, SURREAL '17, MuCo-3DHP '18, JTA '18, CMU-Panoptic '19, GPA '19, 3D People '19, SAIL-VOS '19, HUMBI '20

3DOH50K '20, CHI3D '20, Human4D '20, AIST-Dance++ '21, ASPset-510 '21, IKEA-ASM '21, BML-MoVi '21, Fit3D '21, HumanSC3D '21, HSPACE '21, AGORA '21, SPEC '21, RICH '22, BEHAVE '22

Acknowledgments. This work was supported by the ERC Consolidator Grant project "DeeViSe" (ERC-CoG-2017-773161) and by Robert Bosch GmbH under the project "Context Understanding for Autonomous Systems".