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

NeurIPS 2021

Aligning Silhouette Topology for Self-Adaptive 3D Human Pose Recovery

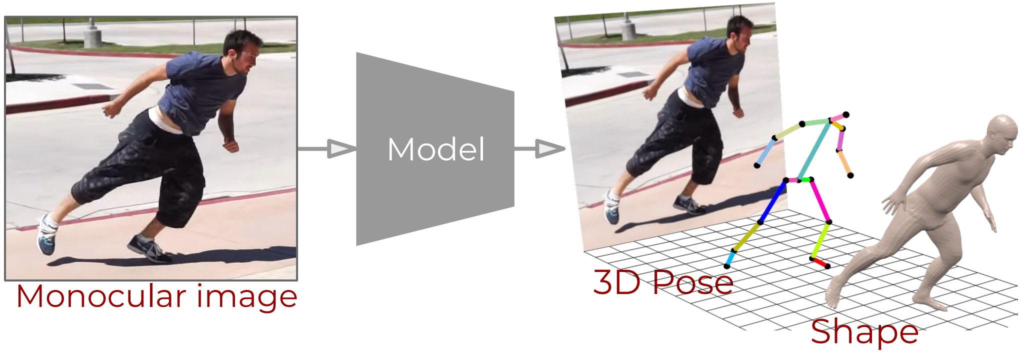
Mugalodi Rakesh*[1] Jogendra Nath Kundu*[1] Varun Jampani[2]
R. Venkatesh Babu[1]

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

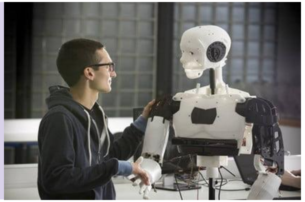
* Equal contribution

Goal task: 3D Human Pose Recovery

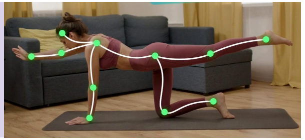
- Inferring the 3D human pose from monocular RGB images.



Human-robot interaction

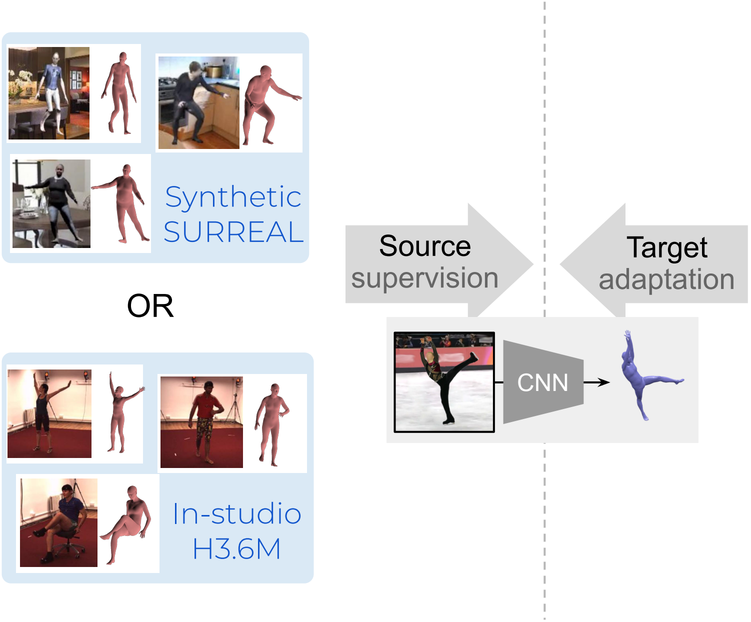


AI fitness tutor



AR/VR application

Domain adaptation: improving deployability of available solution



Mesh recovery requires

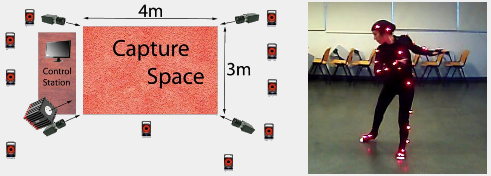
a) Articulation-centric sup.
   (e.g., 3D pose, 2D pose)

b) Shape-centric sup.
   (e.g., Silhouette)

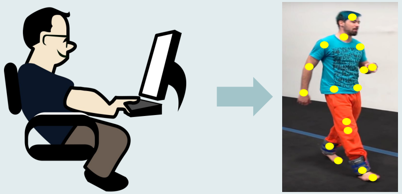
Target Label requirement
(reducing sup. levels)

1. 3D pose GT
2. 2D pose + Silhouette GT
3. Only 2D pose GT
4. Only Silhouette GT

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

Mesh recovery requires

a) Articulation-centric sup. (e.g., 3D pose, 2D pose)

b) Shape-centric sup. (e.g., Silhouette)

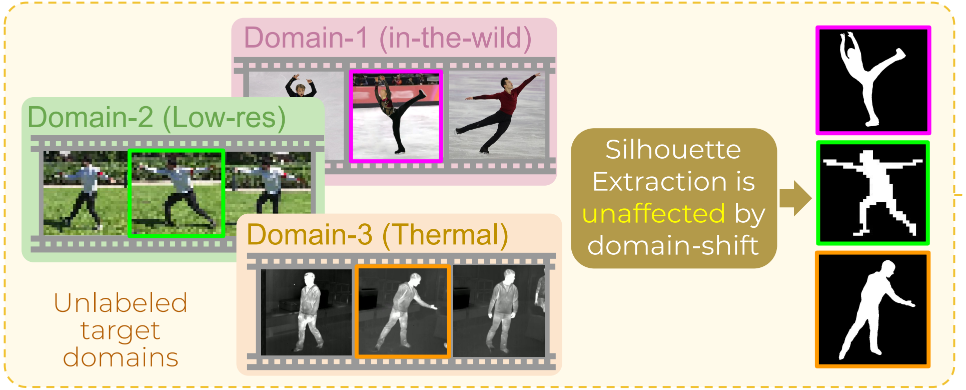
Target Label requirement (reducing sup. levels)

1. 3D pose GT
2. 2D pose + Silhouette GT
3. Only 2D pose GT
4. Only Silhouette GT

One must minimize the target label requirements for convenient deployment.

Domain adaptation: improving deployability of available solution

- Silhouette extracted via classical vision based BG subtraction on static camera feed is found to be considerably robust against domain-shifts.



Mesh recovery requires

a) Articulation-centric sup. (e.g., 3D pose, 2D pose)

b) Shape-centric sup. (e.g., Silhouette)

Target Label requirement (reducing sup. levels)

1. 3D pose GT
2. 2D pose + Silhouette GT
3. Only 2D pose GT
4. Only Silhouette GT ✓

We aim to build an adaptation framework that relies only on silhouette supervision.

Challenges: developing silhouette based self-adaptive framework

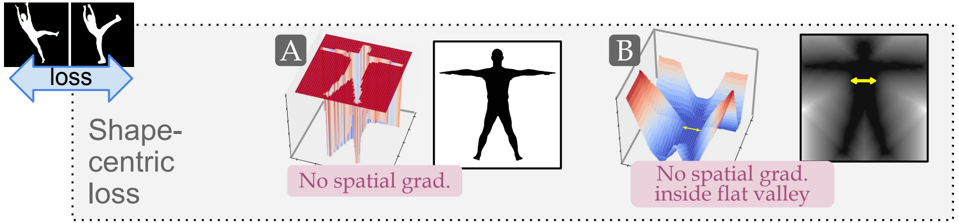
Silhouette losses used in literature:

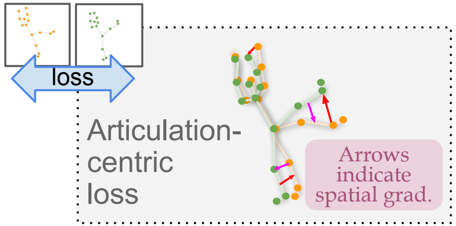
A Pixel-level L1 or cross-entropy → no gradient along spatial direction

B Chamfer loss b/w the 2D silhouette point sets → remains shape-centric

These silhouette losses are not self-sufficient
(Requires to be employed in tandem with a direct 3D or 2D pose supervision)

- Don't provide reliable articulation centric supervision → degenerate solution

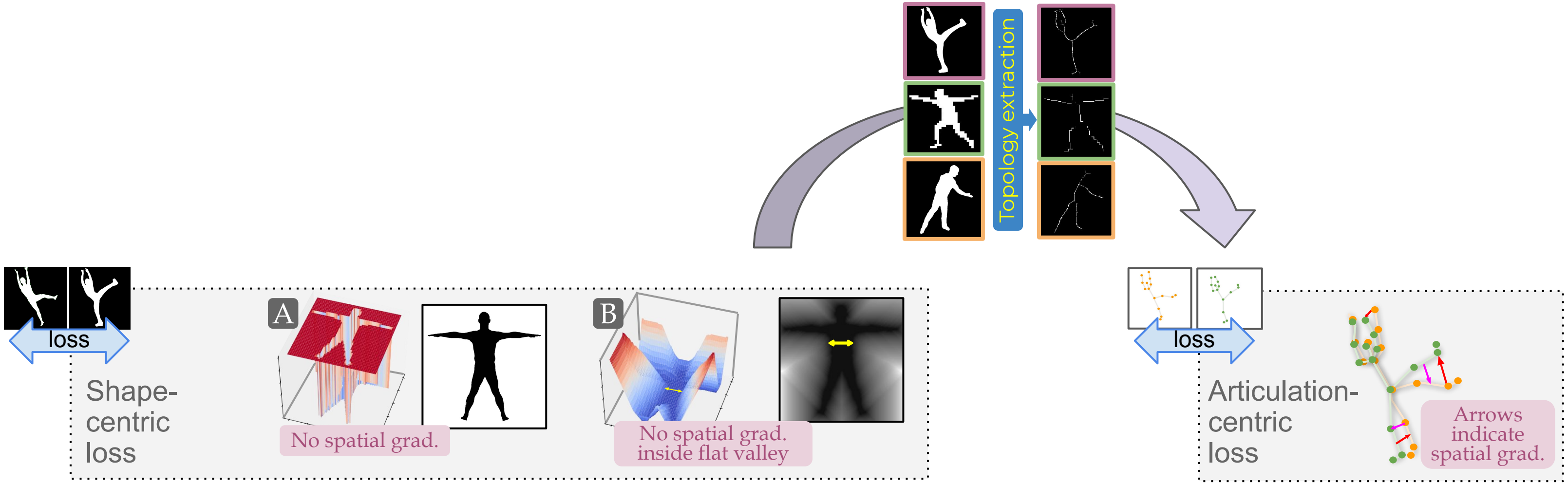




[1] Lassner *et. al.* "Unite the people: Closing the loop between 3d and 2d human representations" , CVPR '17
[2] Pavlakos *et. al.* "Learning to estimate 3d human pose and shape from a single color image. " , CVPR '18

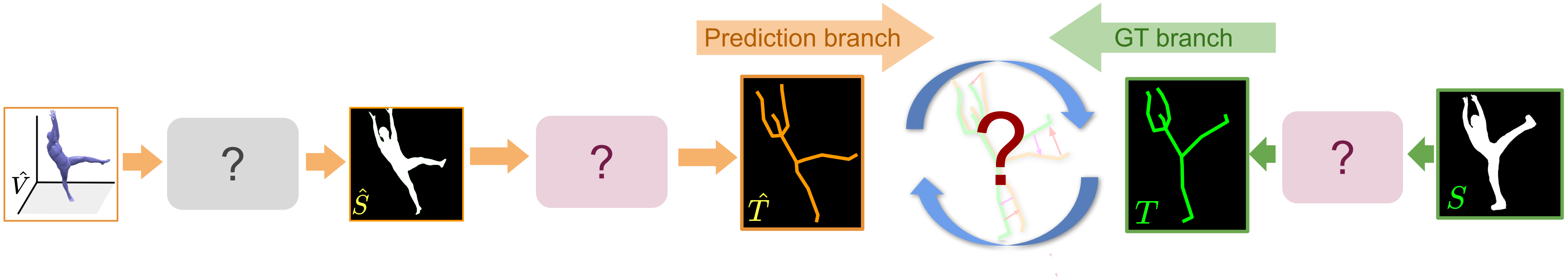
Proposed solution: disentangle topological-skeleton from raw silhouettes

- A new representation, termed as "*topological-skeleton*" to devise a novel self-sufficient silhouette loss.
  (Topological-skeleton is a thin-lined pattern that represents the geometric and structural core of a silhouette mask.)
- This facilitates an auxiliary articulation-centric supervision in the absence of 2D/3D pose GT.



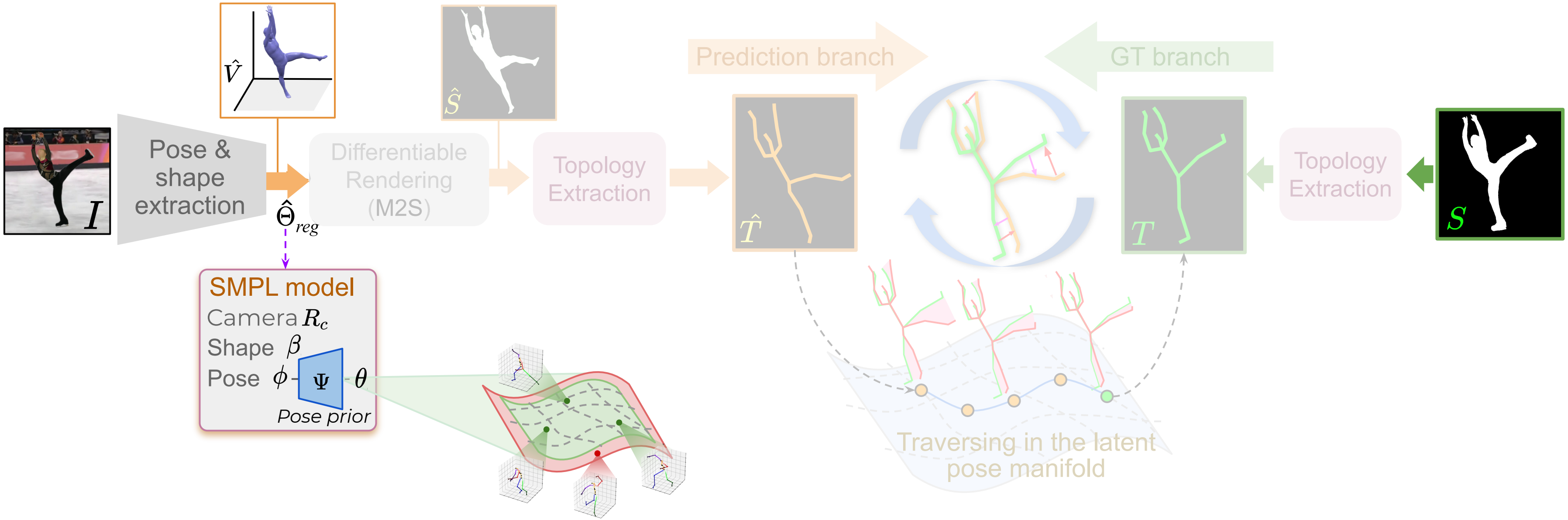
Proposed solution: using topological-skeleton for self-adaptation

- Requirements to realize this framework:
  - a) A way to obtain binary silhouettes from the predicted mesh.
  - b) Differentiable topology extraction module
  - c) A reliable loss on the extracted topology



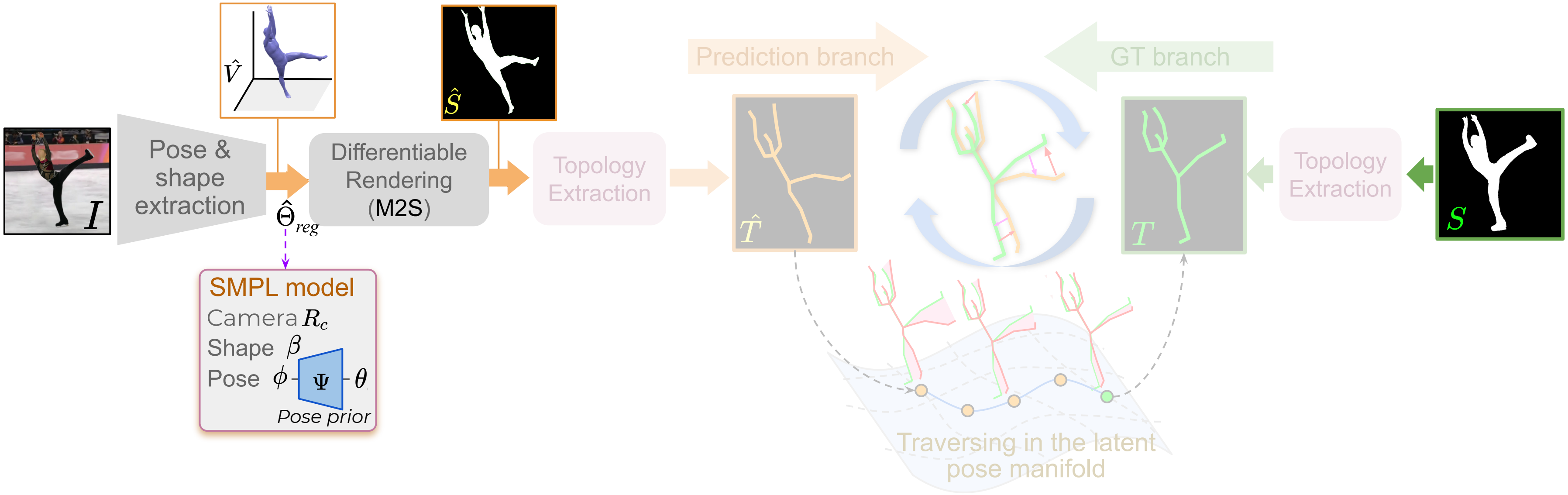
Approach: notations and modules

- Obtaining the mesh for an image **I** via the SMPL regressor.



Approach: notations and modules

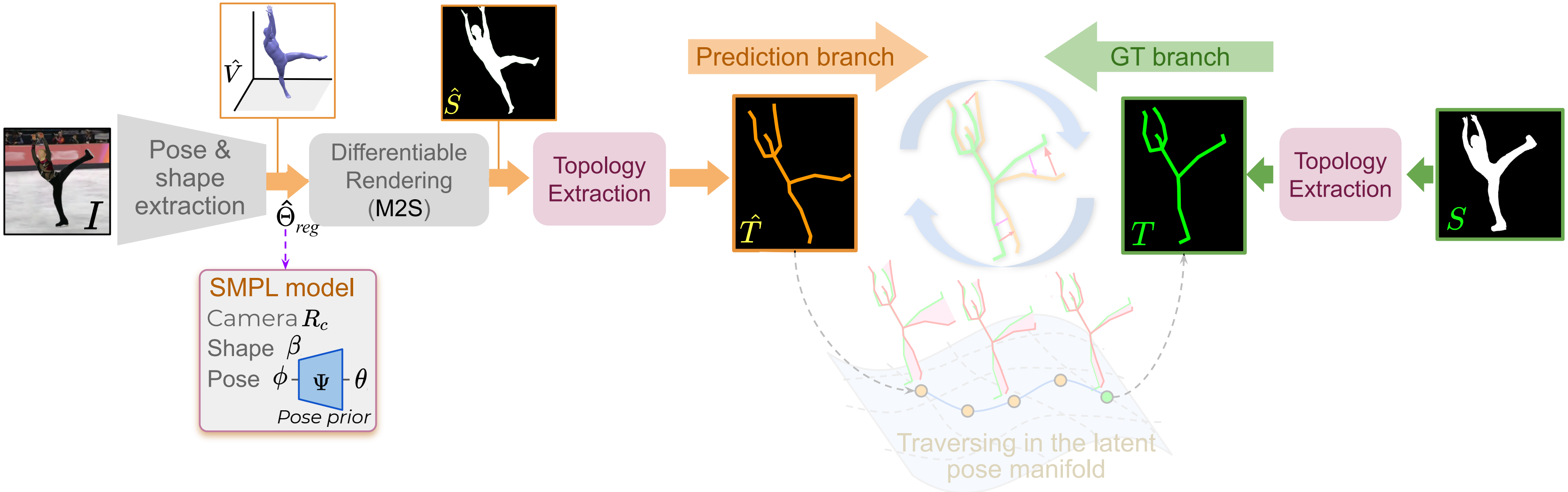
a) M2S: A differentiable rendering module for obtaining silhouettes from predicted mesh.



Approach: notations and modules

a) M2S: A differentiable rendering module for obtaining silhouettes from predicted mesh.

b) A differentiable formulation for extracting topological-skeleton via distance-maps..



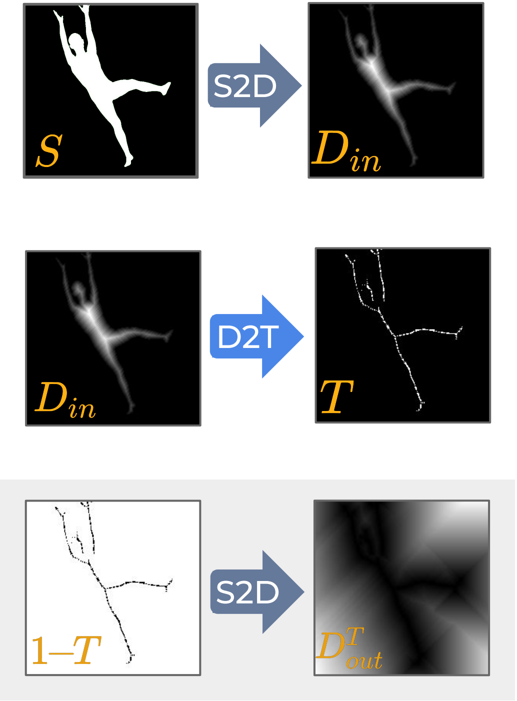
Approach: notations and modules

a) Distance-map, **D** (extracted from silhouette **S**)

- A spatial map D(u), whose intensity at each pixel-location **u** ∈ **U** represents its distance from the closest mask-boundary pixel of **S**.
  - Inwards distance-map, $\mathbf{D_{in}}$(u)
  - Outwards distance-map, $\mathbf{D_{out}}$(u)

b) Topological-skeleton, **T** (extracted from $\mathbf{D_{in}}$)

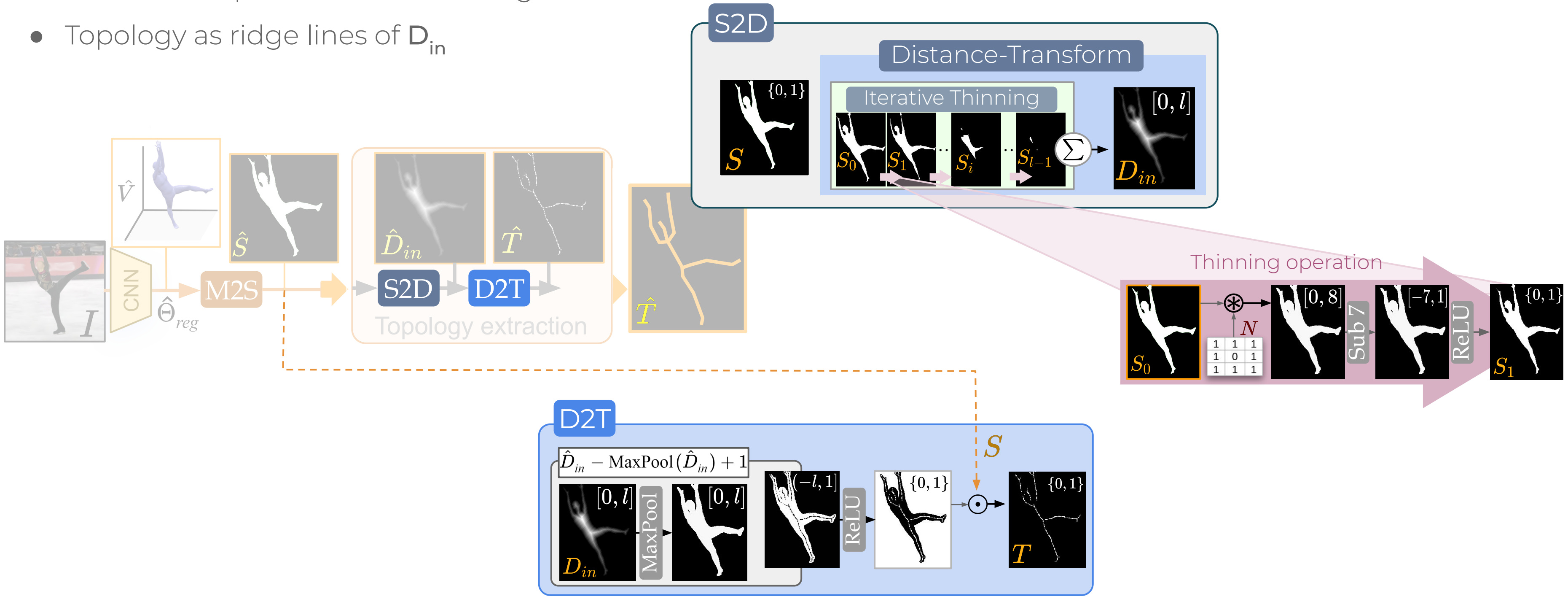
- A thin-lined pattern that represents the geometric and structural core of a silhouette mask **S**.
- Realized as the ridges-lines of $\mathbf{D_{in}}$.



[1] Golland et. al. "Fixed topology skeletons". In CVPR, 2000
[2] Chang et. al. "Extracting skeletons from distance maps". In IJCSNS, 2007.

Approach: internal implementation details of sub-modules

- Distance map via Iterative thinning
- Topology as ridge lines of D_{in}

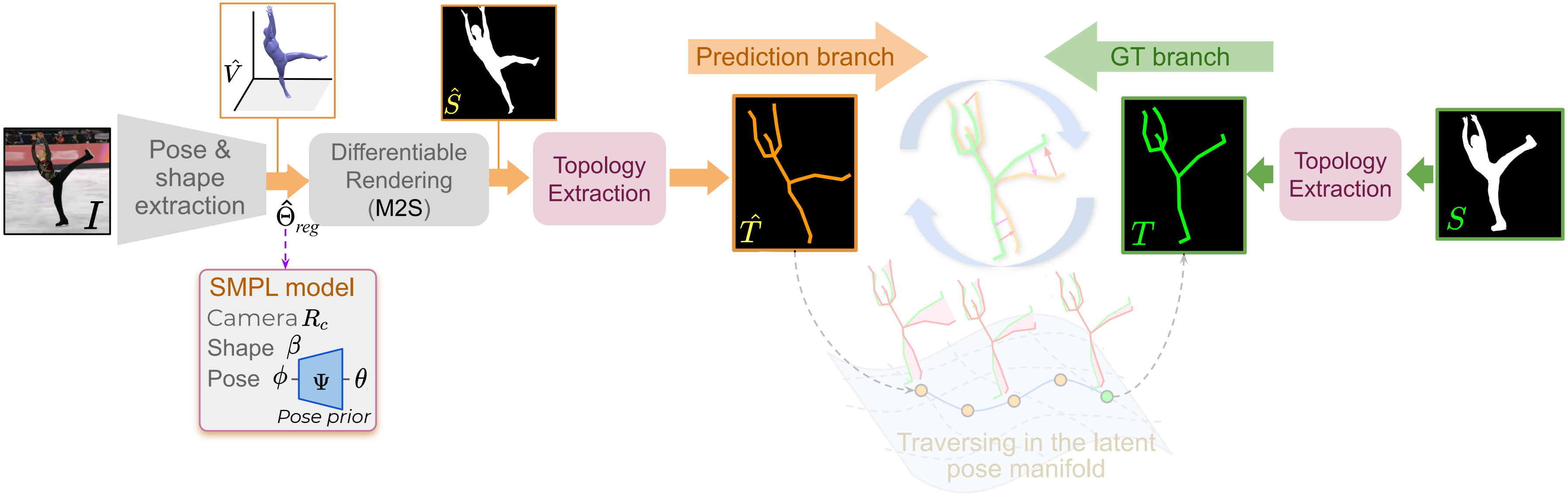


Approach: notations and modules

a) M2S: A differentiable rendering module for obtaining silhouettes from predicted mesh.

b) A differentiable formulation for extracting topological-skeleton via distance-maps..

c) Devising a alignment loss between $\hat{T}$ and T.



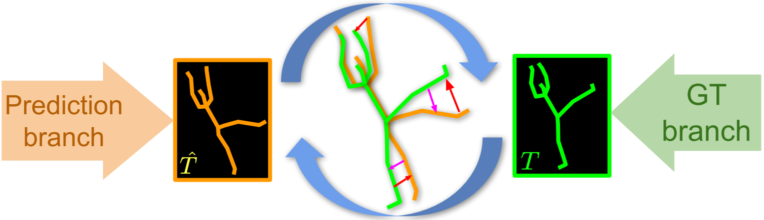
Approach: the topological alignment objective

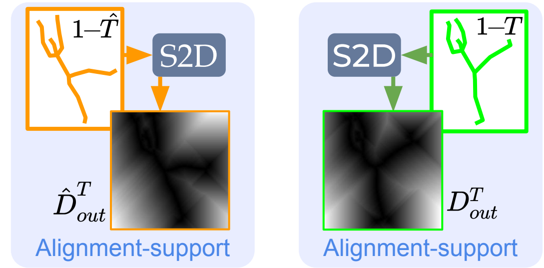
Devising a loss between $\hat{T}$ and T

- L2/L1 loss → no spatial grad.
- Chamfer loss → requires point-set conversion.
- Chamfer inspired loss on spatial maps.

How can we avoid spatial-map to point-set mapping?

Solution: Formalize an equivalent of Chamfer using outwards distance-map D_{out}.



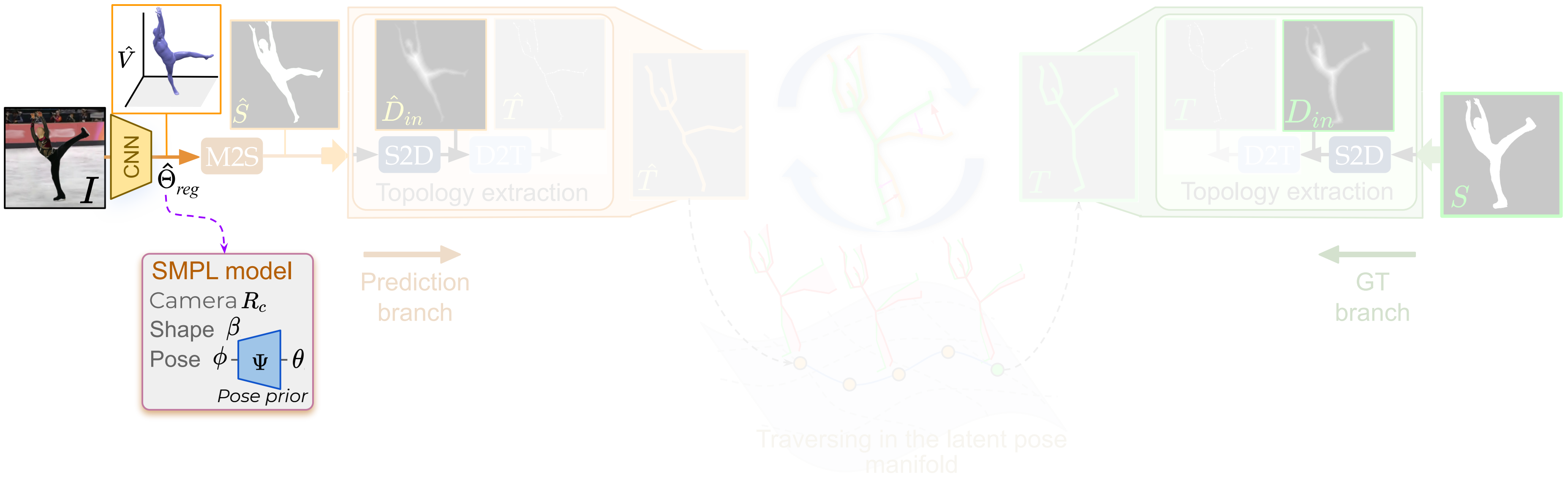


$$\mathcal{L}_T^{(sp)} = \|\hat{D}^T_{out} \odot T\|_{1,1} + \|D^T_{out} \odot \hat{T}\|_{1,1}$$

Approach: a summary

a) silhouette obtained via differentiable rendering module. (M2S)

b) topological-skeleton extracted from silhouettes as ridge lines in distance maps.

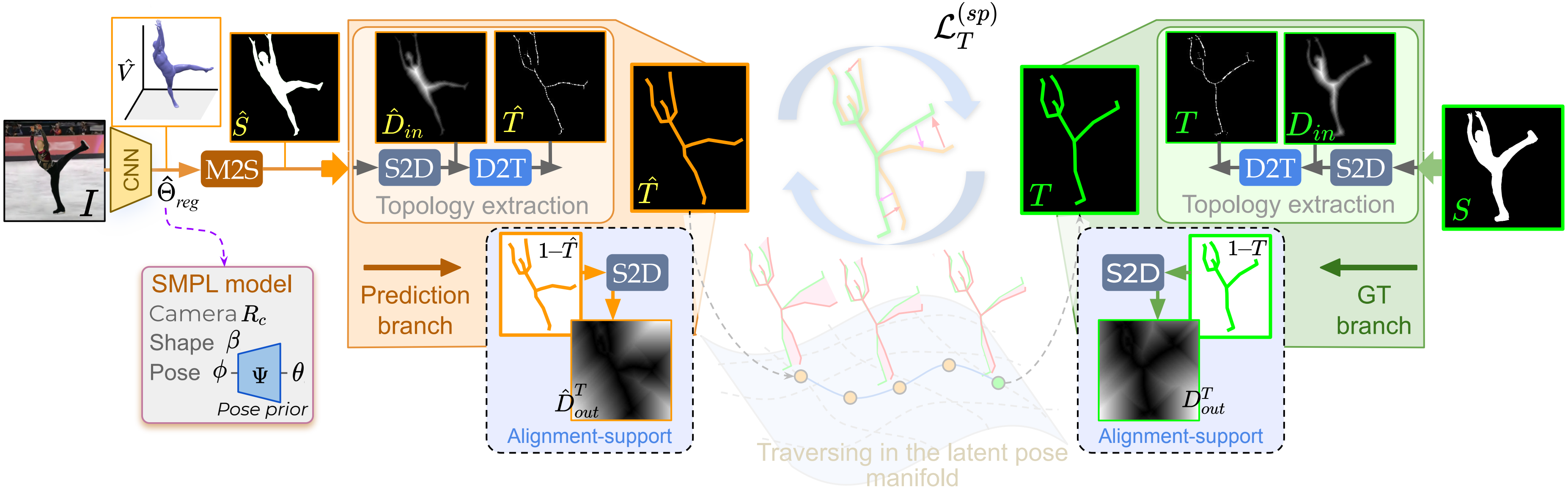


Approach: a summary

a) silhouette obtained via differentiable rendering module. (M2S)

b) topological-skeleton extracted from silhouettes as ridge lines in distance maps.

c) formalized an alignment loss between $\hat{T}$ and T.



Results: adaptation from Synthetic to Real

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

Sup.	Method	PA-MPJPE(↓)
Full	Pavlakos *et al.* [40]	75.9
	HMR [22]	56.8
	SPIN [25]	41.1
Weak	HMR (unpaired) [22]	66.5
	SPIN (unpaired) [25]	62.0
	Ours(S→R,weak)	**58.1**
Unsup.	Kundu *et al.* (unsup) [26]	90.5
	Ours(S→R)	**81.3**

against **self-supervised** prior-arts

Syn. SURREAL → Real H3.6M + MPII → Evaluated on 3DPW

Sup.	Method	MPJPE(↓)	PA-MPJPE(↓)
Full	HMR [22]	128.1	81.3
	Kanazawa *et al.* [23]	116.5	72.6
	SPIN [25]	98.6	59.2
Weak	Martinez *et al.* [34]	-	157.0
	SMPLify [4]	199.2	106.1
	Doersch *et al.* (RGB+2D) [10]	-	82.4
	Ours(S→R, weak)	**126.3**	**79.1**
Unsup.	Doersch *et al.* (DANN) [10]	-	103.0
	Kundu *et al.* (unsup) [26]	187.1	102.7
	Doersch *et al.* (Flow) [10]	-	100.1
	Ours(S→R)	**159.0**	**95.1**

Results: adaptation from Synthetic to Real

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

Syn. SURREAL → Real H3.6M + MPII → Evaluated on 3DPW

against weakly-supervised prior-arts

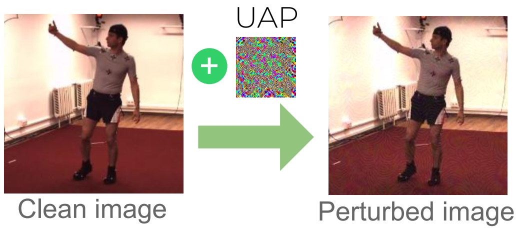
Sup.	Method	PA-MPJPE(↓)
Full	Pavlakos *et al.* [40]	75.9
	HMR [22]	[illegible]
	SPIN [25]	41.1
Weak	HMR (unpaired) [22]	66.5
	SPIN (unpaired) [25]	62.0
	Ours(S→R,weak)	**58.1**
Unsup.	Kundu *et al.* (unsup) [26]	90.5
	Ours(S→R)	**81.3**

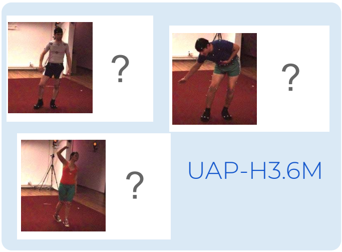
Sup.	Method	MPJPE(↓)	PA-MPJPE(↓)
Full	HMR [22]	128.1	81.3
	Kanazawa *et al.* [23]	116.5	72.6
	SPIN [25]	98.6	59.2
Weak	Martinez *et al.* [34]	-	157.0
	SMPLify [4]	199.2	106.1
	Doersch *et al.* (RGB+2D) [10]	-	82.4
	Ours(S→R, weak)	**126.3**	**79.1**
Unsup.	Doersch *et al.* (DANN) [10]	-	103.0
	Kundu *et al.* (unsup) [26]	187.1	102.7
	Doersch *et al.* (Flow) [10]	-	100.1
	Ours(S→R)	**159.0**	**95.1**

Results: self-adaptation from Real to UAP-H36M

- Universal Adversarial Perturbation (UAP) is an instance-agnostic perturbation that inflict a drop in the task performance.

Real H3.6M + MPII → UAP-H3M → Evaluated on UAP-H3M





	Method	Adaptation from R to UAP-H3M: MPJPE (↓) 4/255	8/255	16/255	PA-MPJPE (↓) 4/255	8/255	16/255
Pre-Adapt.	SPIN [25]	65.8	98.2	160.1	44.6	60.8	90.7
	Ours(R)	67.7	103.9	161.8	46.9	63.6	91.2
Post-Adapt.	A1: SPIN+$\mathcal{L}_{2D}^{(p)}$	64.5	94.0	151.2	43.4	59.5	89.8
	A2: SPIN+$\mathcal{L}_{2D}^{(p)}$+$\mathcal{L}_{S}^{(p)}$	64.1	89.1	136.5	43.4	58.9	85.1
	Ours(R→UAP)	**63.6**	**84.7**	**125.2**	**43.2**	**57.6**	**79.4**

[1] Moosavi-Dezfooli *et al.*, "Universal adversarial perturbations.", CVPR '17

Results: self-adaptation from Real to LR-3DPW

- Low resolution (LR) images inflict a drop in task performance.

Normal image

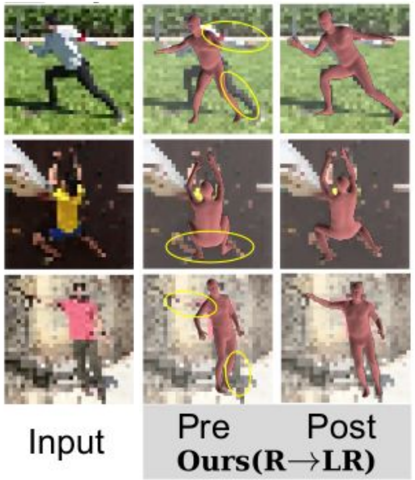
Low-res image

Real H3.6M + MPII → LR-3DPW → Evaluated on LR-3DPW

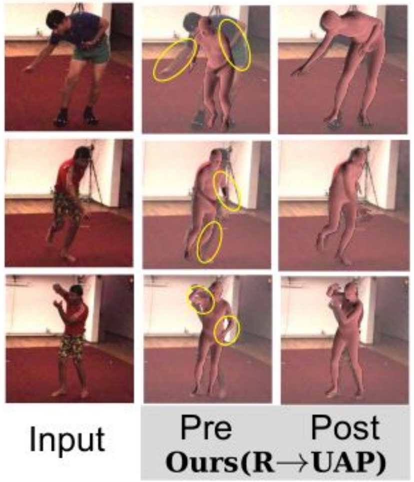
	Method	Adaptation from R to LR-3DPW					
		MPJPE (↓)			PA-MPJPE (↓)		
		96	52	32	96	52	32
Pre-Adapt.	SPIN [25]	104.3	120.3	176.4	63.7	71.1	87.9
	Ours(R)	110.8	127.5	178.1	68.6	76.3	88.2
Post-Adapt.	A1: SPIN+$\mathcal{L}_{2D}^{(p)}$	100.2	117.0	153.6	61.7	70.3	85.4
	A2: SPIN+$\mathcal{L}_{2D}^{(p)}$+$\mathcal{L}_{S}^{(p)}$	100.1	115.2	147.5	61.5	69.8	82.3
	Ours(R→LR)	**99.8**	**114.7**	**134.2**	**61.3**	**68.7**	**78.9**

Results: qualitative comparison

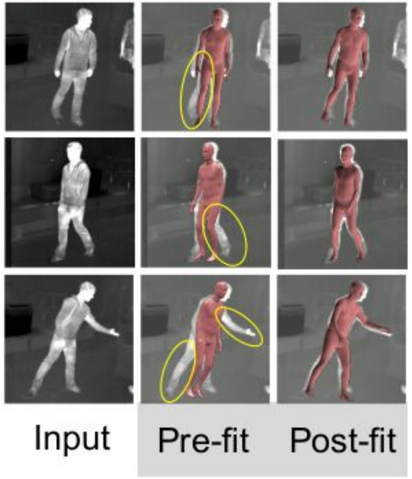
Real H3.6M + MPII → LR-3DPW



Real H3.6M + MPII → UAP-H3M

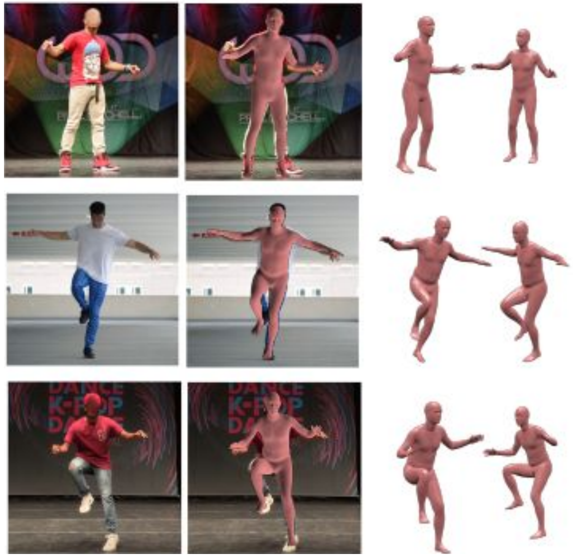


Real H3.6M + MPII Thermal



Qualitative results: adaptation from Synthetic to Real

Real
Internet



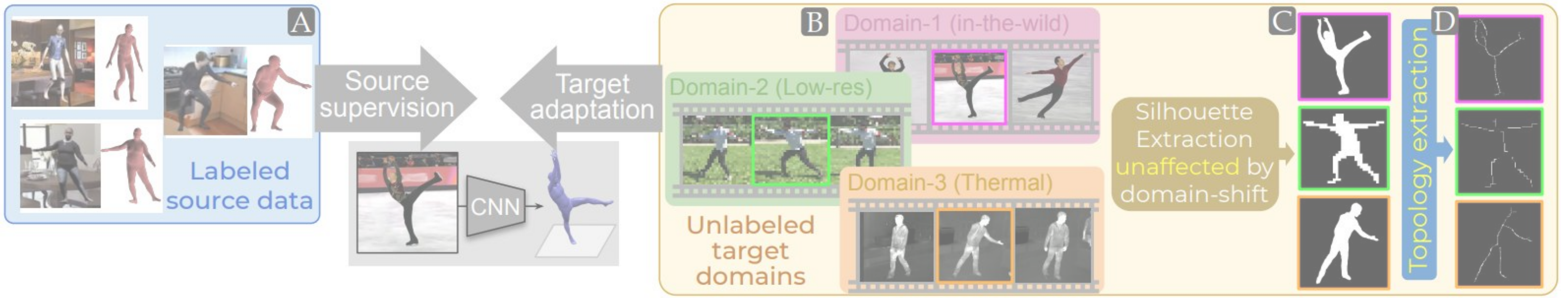
Real
H3.6M

Real
3DPW

Summary

- We propose a self-supervised domain adaptation framework that relies only on silhouette supervision.
- We develop a series of convolution-friendly and differentiable spatial transformations in order to disentangle a topological-skeleton representation from raw silhouettes.
- We devise a Chamfer-inspired spatial alignment loss via distance map computation, effectively avoiding any gradient hindering spatial-to-pointset conversion.

A step towards next generation deployment friendly (i.e. self-adaptive) human mesh recovery systems.



Thank You!

Aligning Silhouette Topology for Self-Adaptive 3D Human Pose Recovery

Please check our project page for more details

https://sites.google.com/view/align-topo-human