Mask as Supervision: Leveraging Unified Mask Information for Unsupervised 3D Pose Estimation

Yuchen Yang[1,2,*] Yu Qiao[2] Xiao Sun[2]

[1]Fudan University [2]Shanghai Artificial Intelligence Laboratory

⋆ Work performed during his internship at Shanghai Artificial Intelligence Laboratory.

Introduction

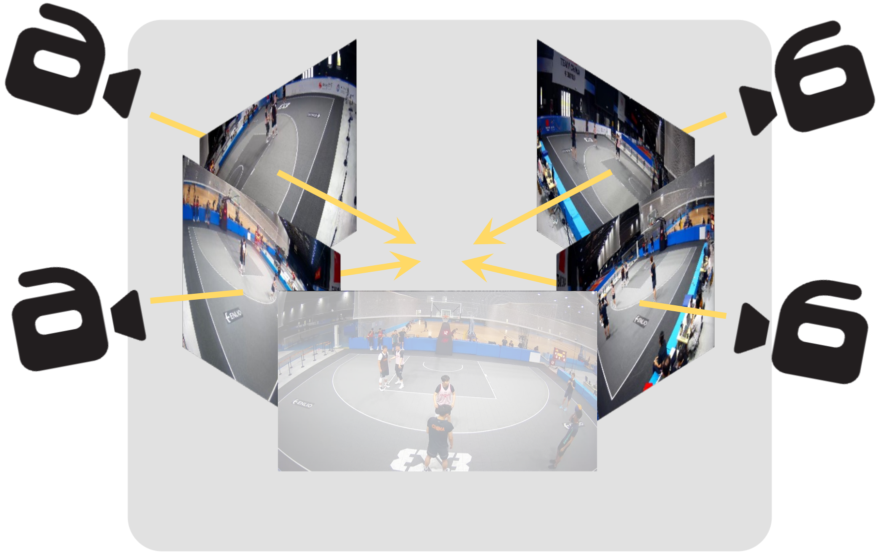
➢ Why unsupervised pose estimation?

Problem Definition – 3D HPE

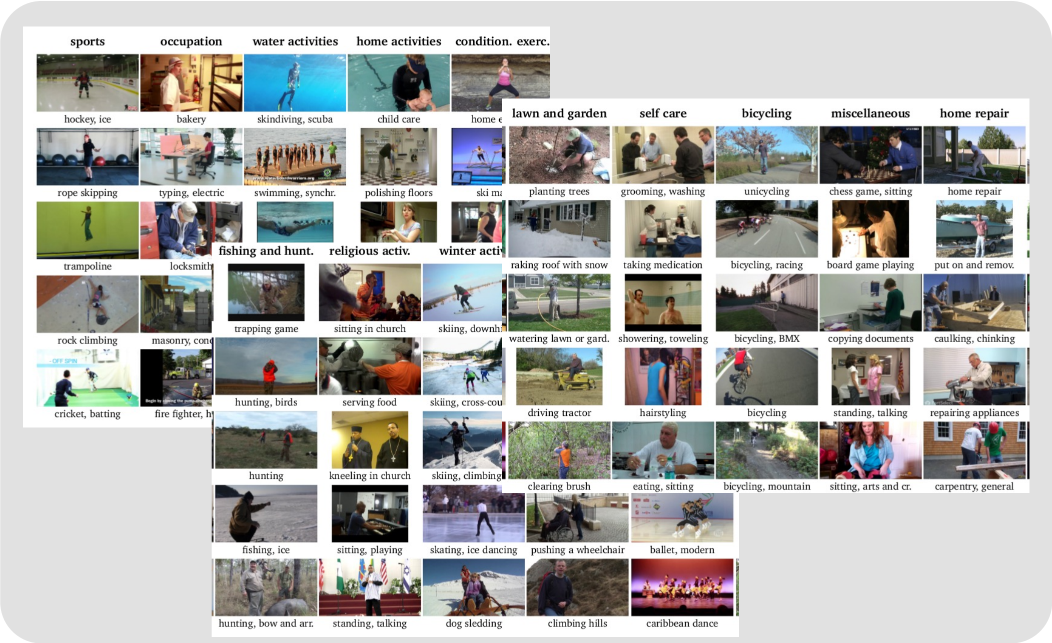
$$X = \phi(I)$$

Given an input image I, determine a set of joints $X \in \mathbb{R}^{J\times 3}$.
J represents the number of joints.

- annotate 3D data remains costly
- vast in-the-wild data for generalization



Motion Capture Environment

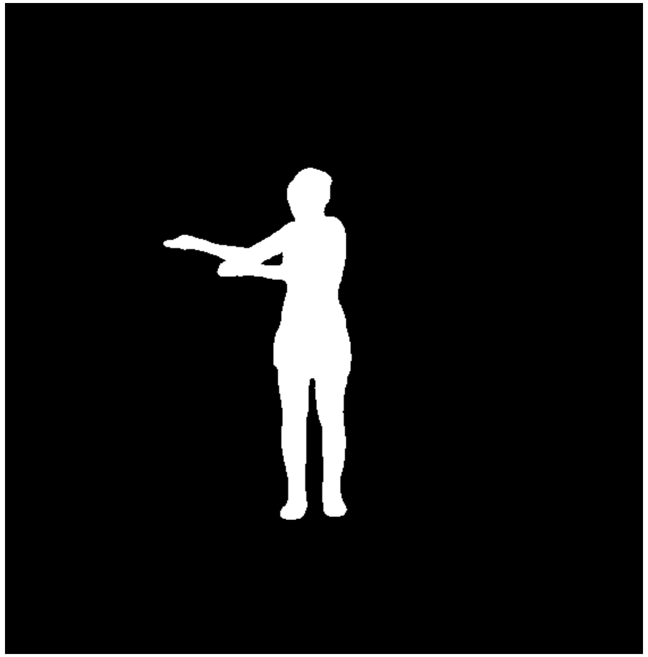


In-the-wild

* Mykhaylo Andriluka, et al. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. CVPR2024.

➢ Why mask as supervision?

For human, we can easily estimate keypoints from masks

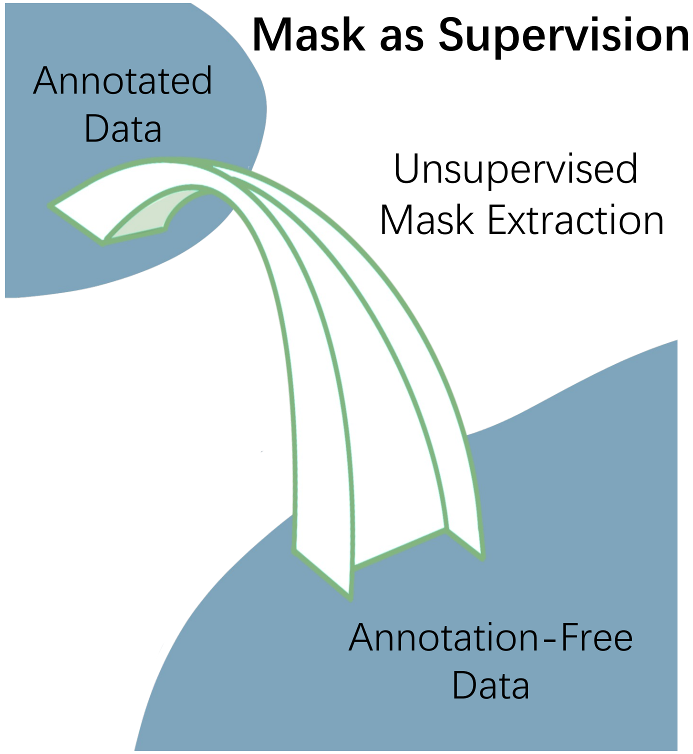


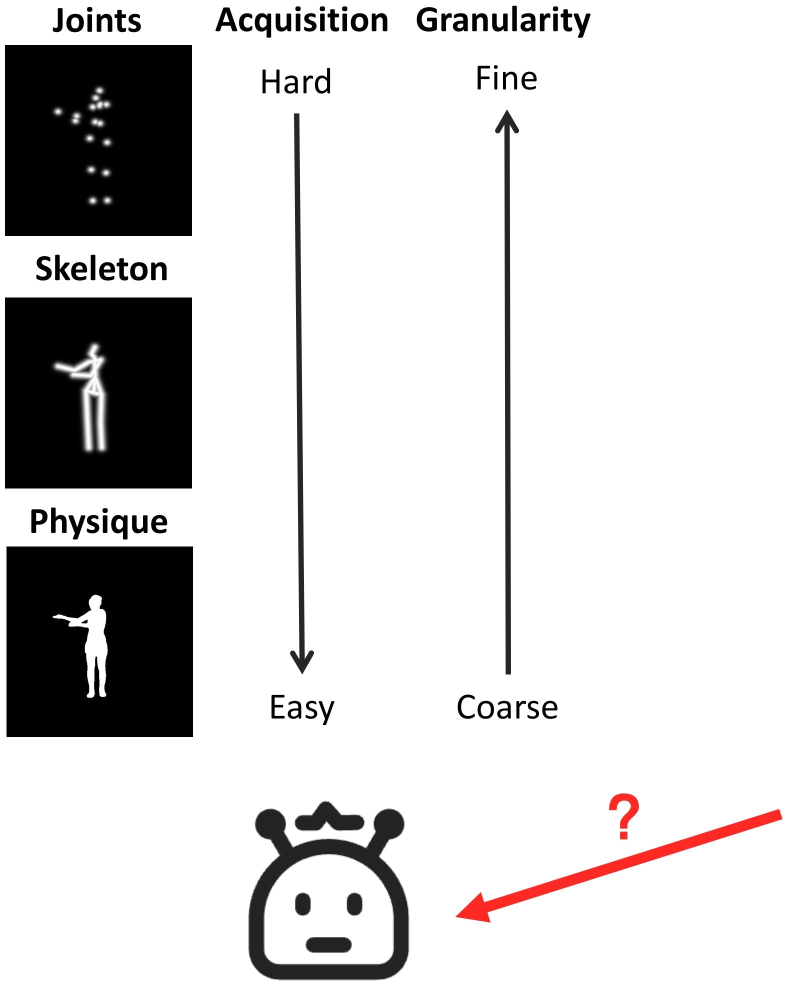
- Rich priors embedded in mask
- Human priors already acquired

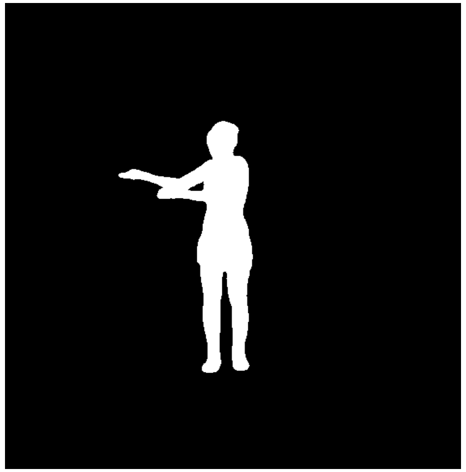
?

Introduction

- Why mask as supervision?

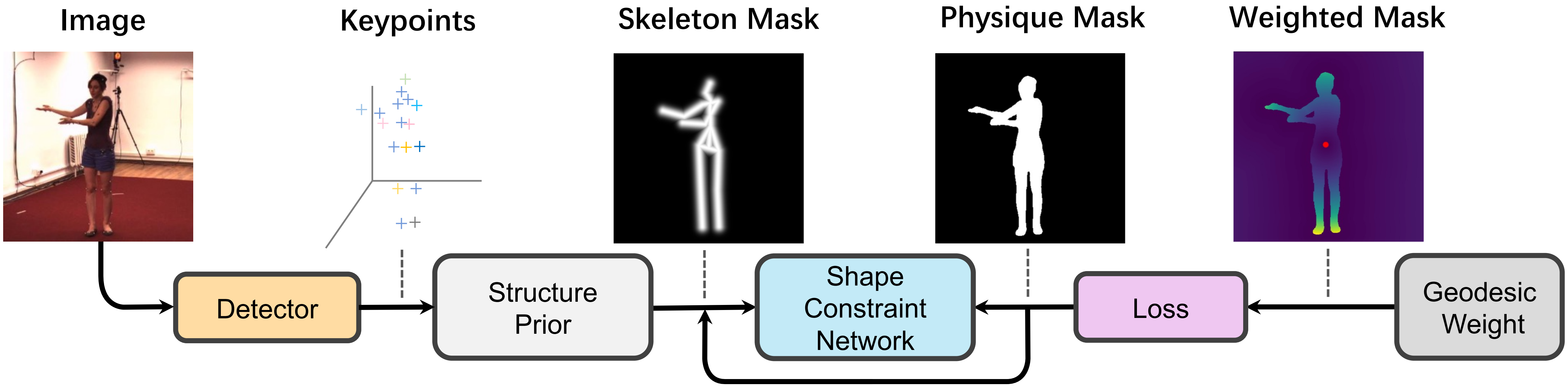






Method

- ➢Coarse-to-fine framework
- ➢Structure prior and shape constraint



Method

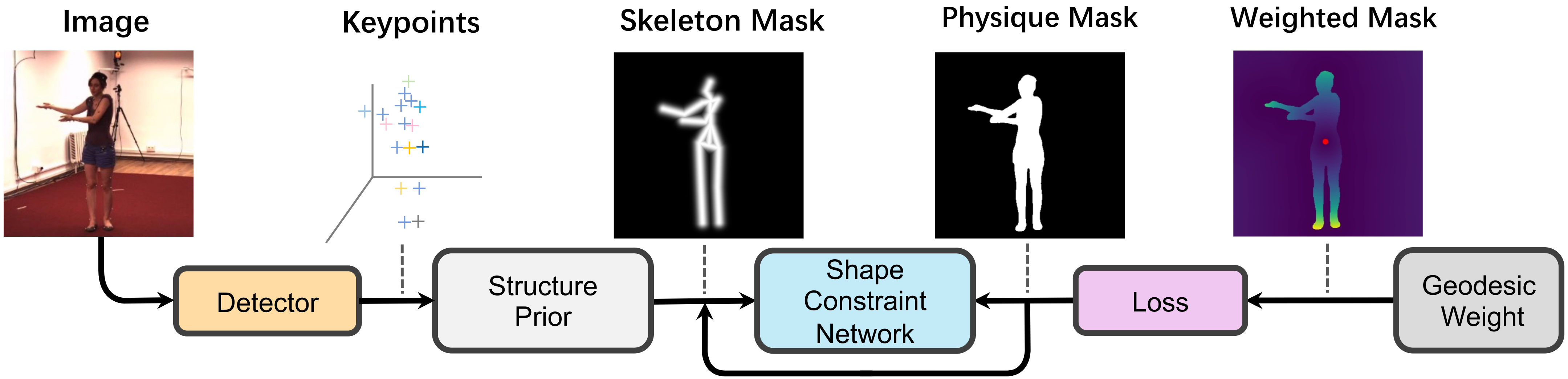
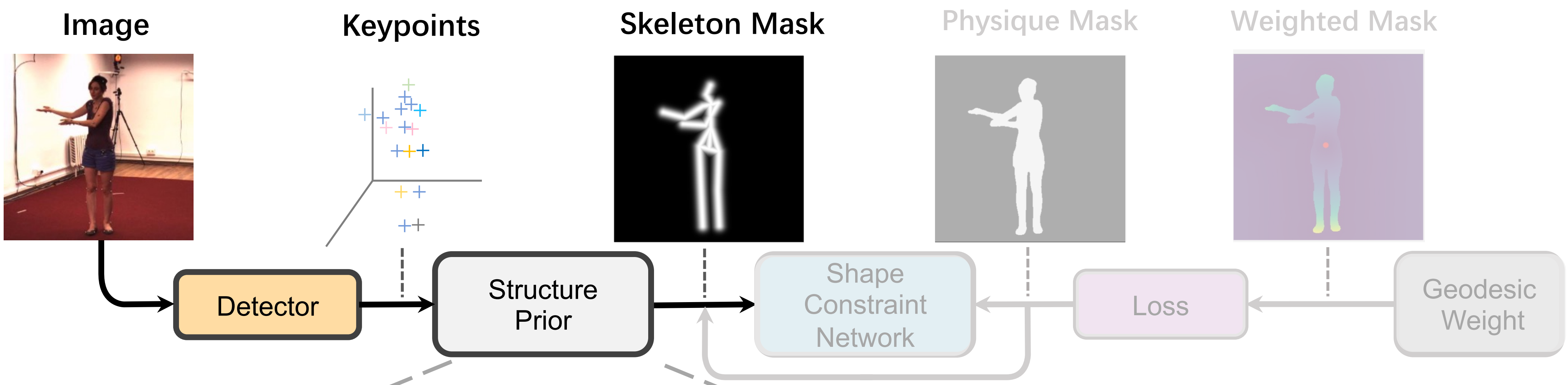
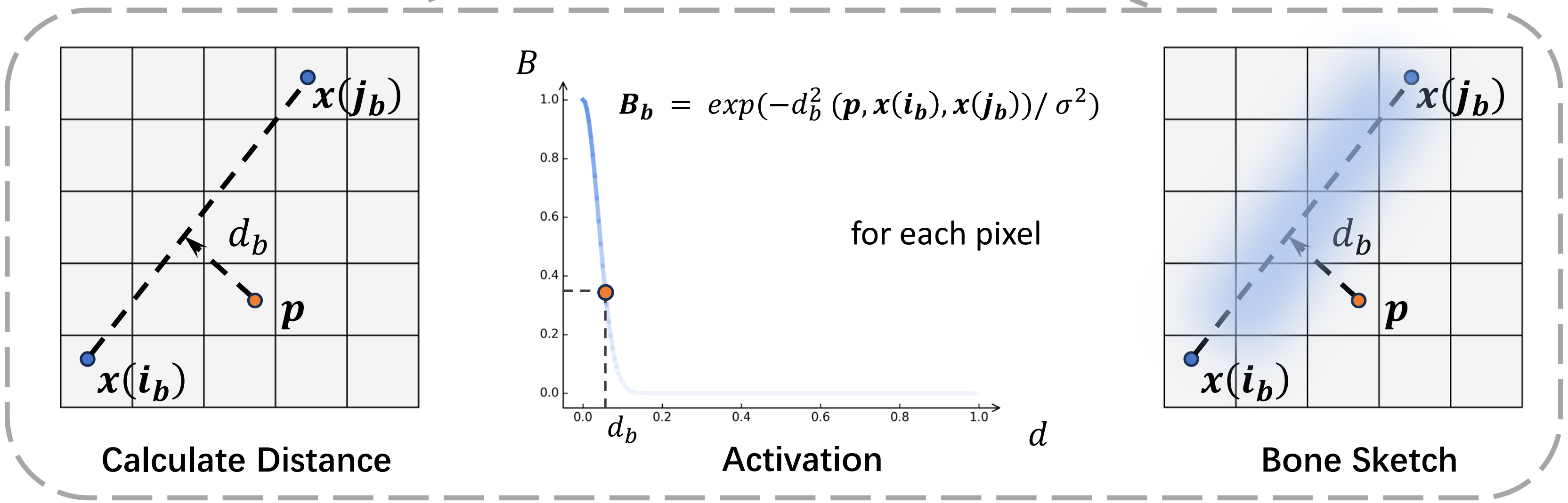


Image
Keypoints
Skeleton Mask
Physique Mask
Weighted Mask
Detector
Structure Prior
Shape Constraint Network
Loss
Geodesic Weight

Method





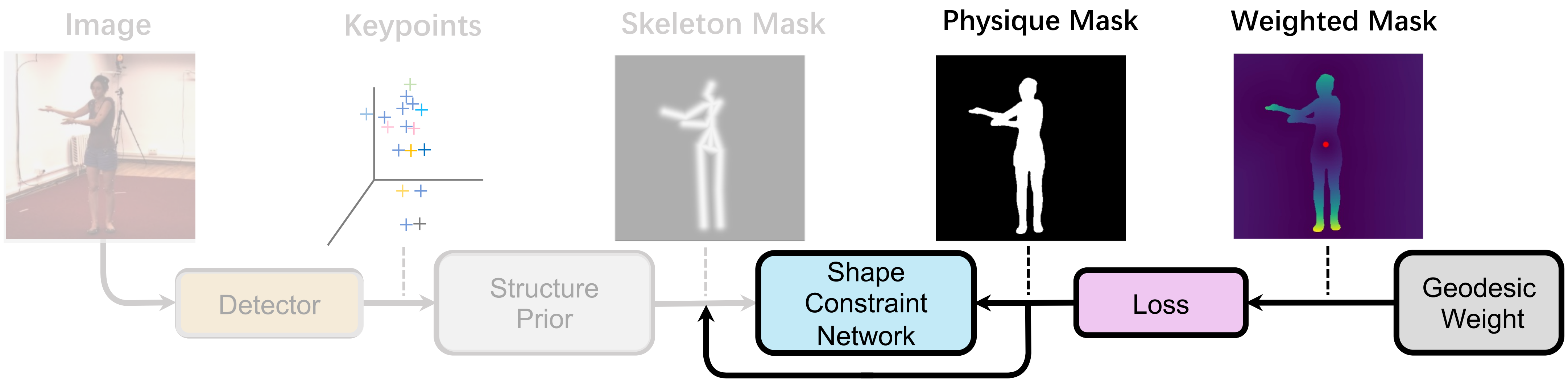
Skeleton Mask

$$B_b = exp(-d_b^2(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{i_b}), \boldsymbol{x}(\boldsymbol{j_b}))/\sigma^2)$$

$$M_{Skel}(x) = \sum_{b=1}^{B} B_b(x)$$

σ is a hyper-parameter controlling bone width.
The line segment defined by bone (x_1, x_2).
Skeleton Mask M_{skel} is from all bone maps via pixel-wise summation.

Method



Physique Mask

$$M_{\text{Physo}}(x) = \psi(M_{Skel}(x))$$

ψ is implemented by a U-Net ended with a sigmoid function to make $M_{\text{Physo}} \in (0,1)$

Loss Function

$$\mathcal{L} = \lambda_s \left|\left|G \odot \left(M^{gt} - M_{Skel}(x)\right)\right|\right|_2^2 + \lambda_p \left|\left|G \odot \left(M^{gt} - M_{Phso}(x)\right)\right|\right|_2^2$$

Geodesic Distance, denoted as G, can be computed using the fast marching method

with mask centroid as zero point.

M^{gt} is the given ground truth mask as supervision.

λ is the balancing factor for loss.

Experiments

SPP-based Method

Step1: Predict landmarks $\mathbb{R}^{L\times3} (L \geq 2 \times J)$

Step2: Train a mapping network θ: $\mathbb{R}^{L\times3} \rightarrow \mathbb{R}^{J\times3}$

- Ignore left-right reversal problem

- Involve human annotation

Quantitative Results

Table 2: Comparison with state-of-the-art methods on MPI-INF-3DHP. MPJPE is in *cm*. Note that the first four methods use supervised post-processing and Sosa *et al.* [45] uses unpaired 2D pose to obtain interpretable keypoints.

Method	PCK(↑)	AUC(↑)	MPJPE(↓)
Denton *et al.* [6]	-	-	22.28
Rhodin *et al.* [38]	-	-	20.24
Honari *et al.* [14]	-	-	20.95
Honari *et al.* [15]	-	-	14.57
Sosa *et al.* [45]	69.6	32.8	-
Ours	60.2	24.7	19.36
Ours (SPP)	**71.3**	**42.7**	**13.67**

Table 1: Comparison with state-of-the-art methods on Human3.6M. **SPP**: supervised post-processing. **UP**: unpaired ground truth pose or its prior, **T**: manually designed template. **SF**: supervised flip to eliminate left-right ambiguity. † indicates our results do not consider the ambiguity in left-right reversal. †† indicates we do not consider inner skeleton relationships and follow the common SPP settings. The best results in SPP and No SPP groups are marked in red and blue. MSE is in 2D % and MPJPEs are in *mm*.

	Method	Settings				Metrics (↓)			
		UP	T	SF	Joint	MSE	MPJPE	N-MPJPE	P-MPJPE
SPP	Thewlis *et al.* [50]	✗	✗	✓	2D	7.51	-	-	-
SPP	Zhang *et al.* [57]	✗	✗	✓	2D	4.14	-	-	-
SPP	Lorenz *et al.* [30]	✗	✗	✓	2D	2.79	-	-	-
SPP	Suwajanakorn *et al.* [49]	✗	✗	✗	3D	-	158.7	156.8	112.9
SPP	Sun *et al.* [47]	✗	✗	✗	3D	-	125.0	-	105.0
SPP	Honari *et al.* [14]	✗	✗	✓	3D	-	100.3	99.3	74.9
SPP	Honari *et al.* [15]	✗	✗	✗	3D	**2.38**	73.8	72.6	63.0
SPP	**Ours**††	✗	✗	✗	3D	2.52	**65.5**	**66.1**	**61.9**
No SPP	Schmidtke *et al.* [40]	✗	✓	✓	2D	3.31	-	-	-
No SPP	Jakab *et al.* [20]	✓	✗	✓	2D	**2.73**	-	-	-
No SPP	Sosa *et al.* [45]	✓	✗	✓	3D	-	-	-	96.4
No SPP	Kundu *et al.* [23]	✓	✓	✓	3D	-	99.2	-	-
No SPP	Kundu *et al.* [24]	✓	✗	✓	3D	-	-	-	89.4
No SPP	**Ours**†	✗	✗	✓	3D	3.17	**85.6**	**85.6**	**79.3**
No SPP	**Ours**	✗	✗	✗	3D	3.63	95.9	96.8	90.4

Experiments

Quantitative Results

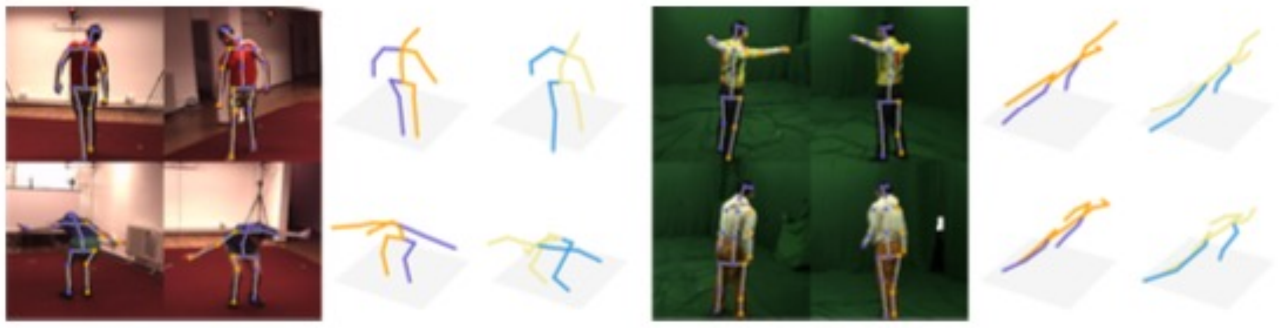


Table 3: Ablation study on shape reconstruction.

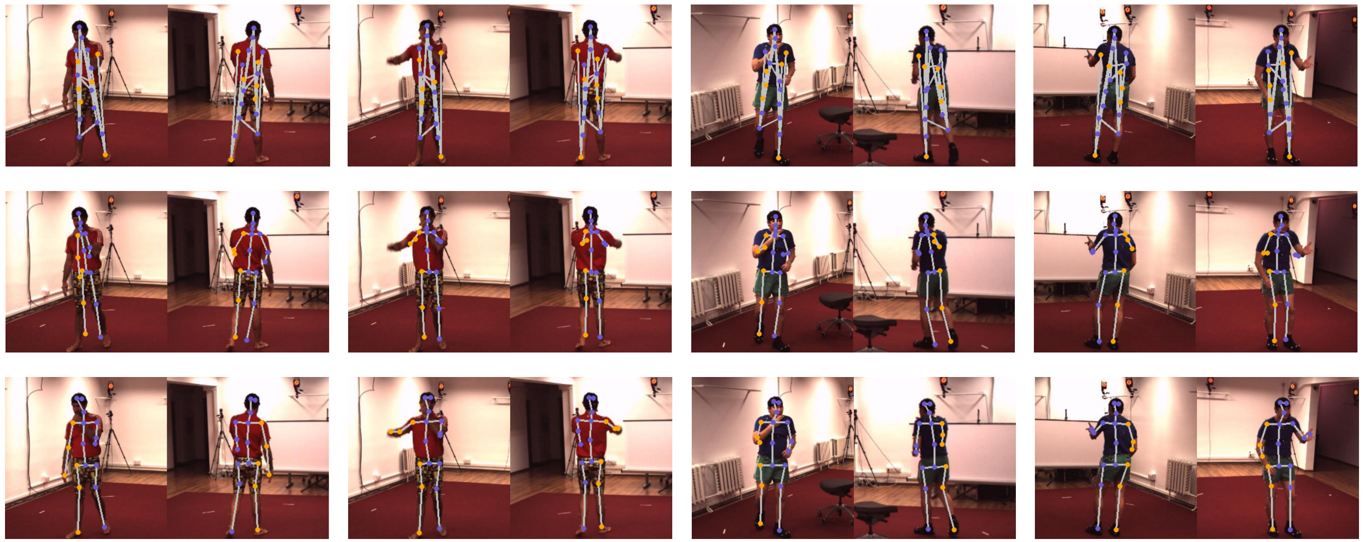
Configurations	MPJPE (↓)	Ambiguity Ratio (↓)
wo ψ, **G**, Δ	118.1	48.73%
wo ψ, **G**	127.4	23.34%
wo **G**	102.6	22.83%
Full	95.9	20.33%

Ablation Study

w/o Structural Prior

(supervised) Post Processing

(unsupervised) Ours

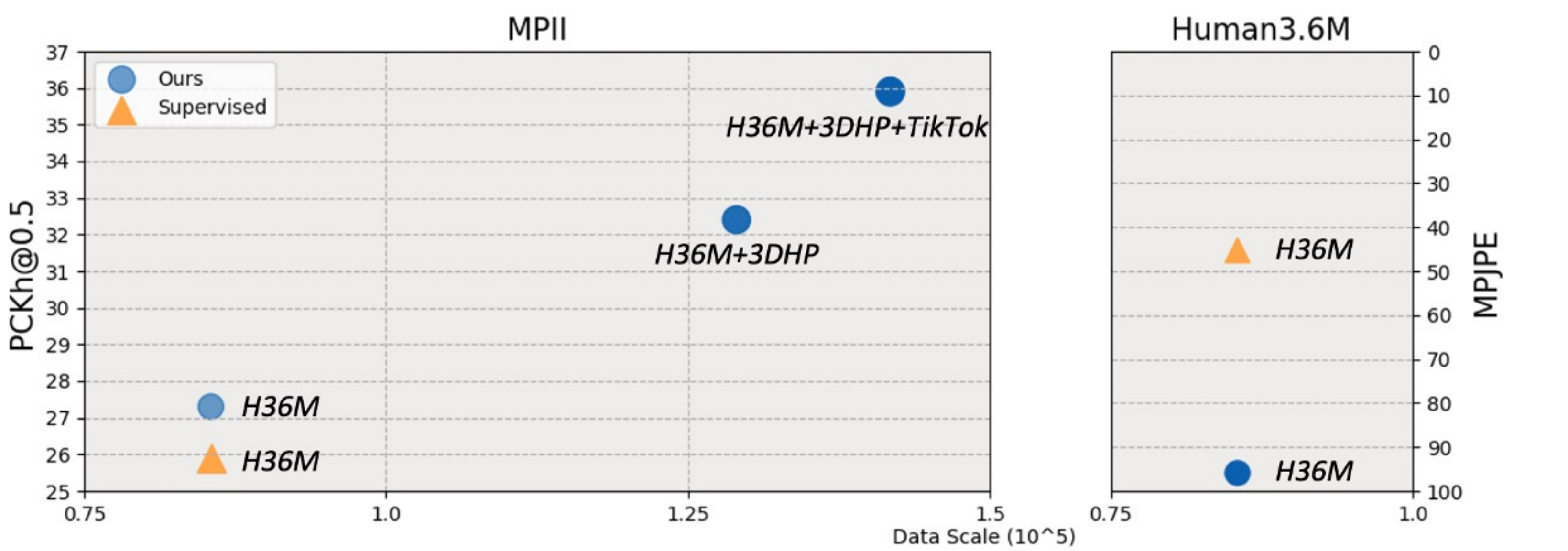


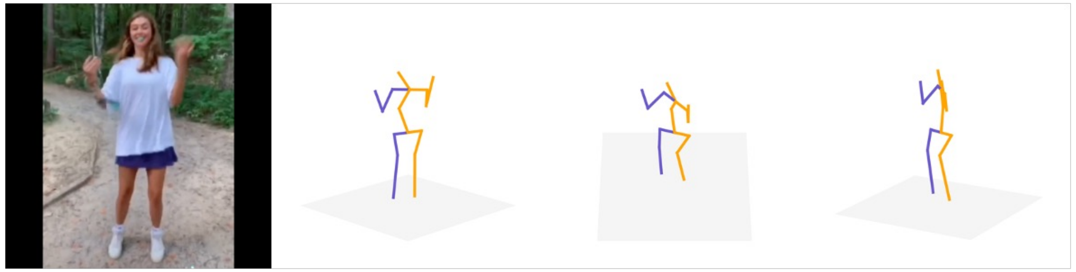
- Effectiveness of Skeleton Mask

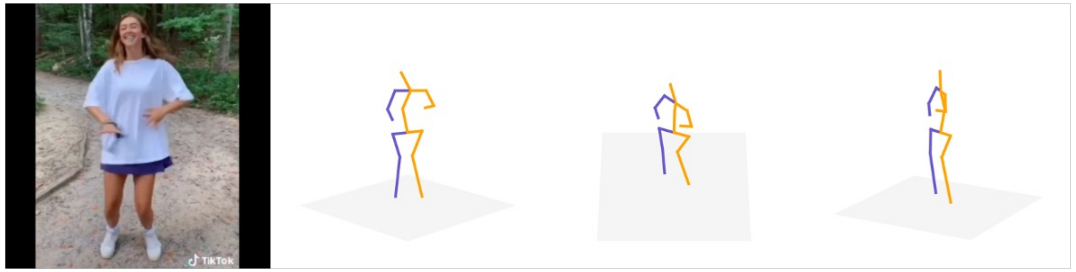
- Effectiveness of Physique Mask and the rest components

Experiments

Leveraging In-the-wild Data







Thanks!