CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation

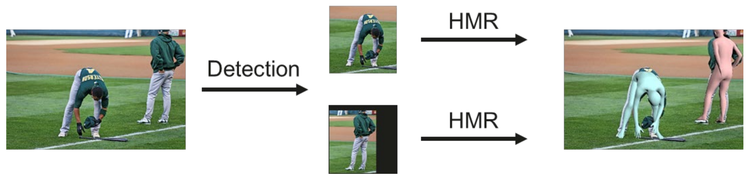
Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan
Huawei Noah's Ark Lab

ECCV TEL AVIV 2022

1. Introduction

Task: reconstruct human body meshes in an image

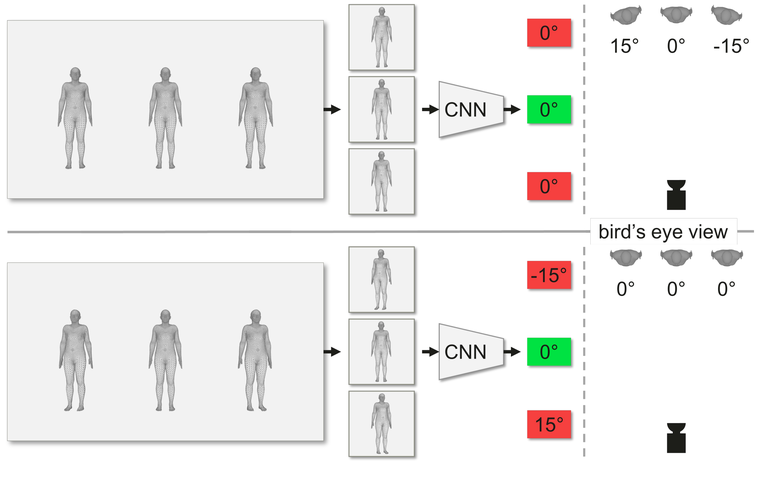
- Top-Down Approach (dominating)



- Bottom-up Approach

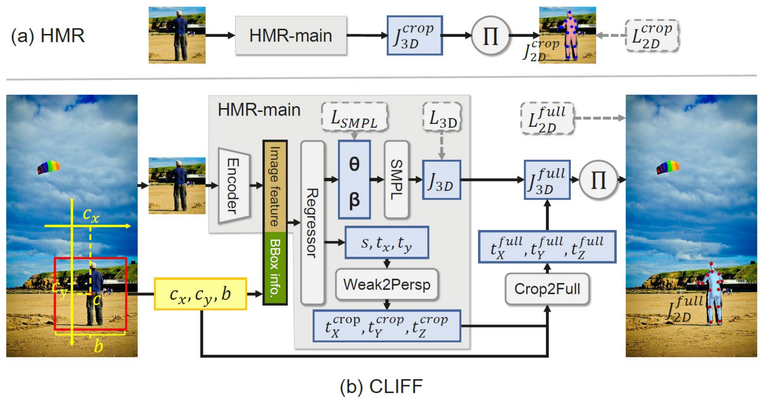
2. Motivation

- Cropping discards location information, and causes inaccurate global rotation estimation.
- Pseudo-GT of in-the-wild images helps regression-based models a lot.



3. Approach

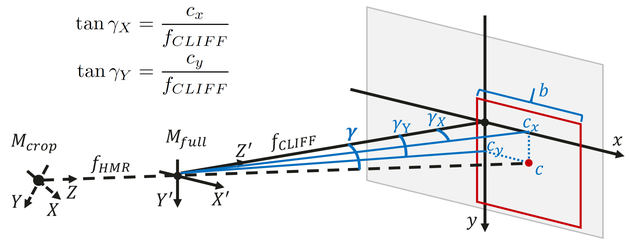
- Take HMR as baseline, and make two modifications



1. Additional input: bounding box information

$$I_{bbox} = [\frac{c_x}{f_{CLIFF}}, \frac{c_y}{f_{CLIFF}}, \frac{b}{f_{CLIFF}}]$$

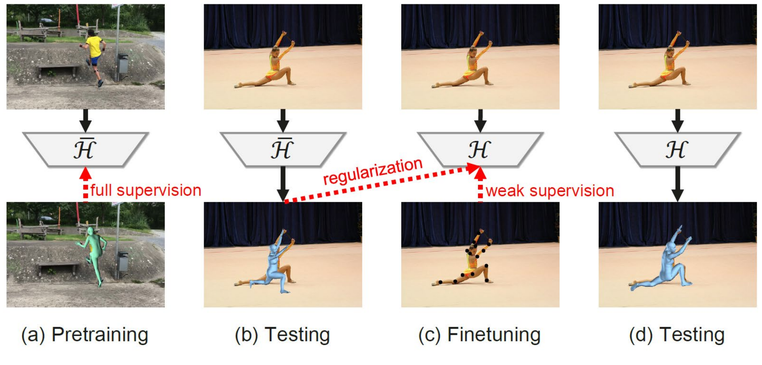
$$\tan\gamma_X = \frac{c_x}{f_{CLIFF}} \qquad \tan\gamma_Y = \frac{c_y}{f_{CLIFF}}$$



2. 2D Reprojection loss in the full image

$$t_X^{full} = t_X^{crop} + \frac{2 \cdot c_x}{b \cdot s}, \quad t_Y^{full} = t_Y^{crop} + \frac{2 \cdot c_y}{b \cdot s}, \quad t_Z^{full} = t_Z^{crop} \cdot \frac{f_{CLIFF}}{f_{HMR}} \cdot \frac{r}{b}$$

- CLIFF+: CLIFF-based pseudo-GT annotator



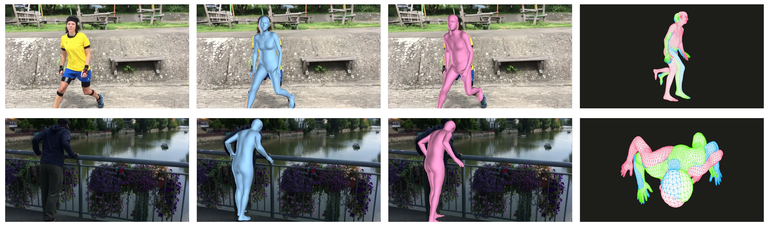
4. Experiments

- CLIFF outperforms SOTA by significant margins

Table 1. Performance comparison between CLIFF and state-of-the-art methods on 3DPW, Human3.6M and AGORA

	Method	3DPW MPJPE ↓	3DPW PA-MPJPE ↓	3DPW PVE ↓	Human3.6M MPJPE ↓	Human3.6M PA-MPJPE ↓	AGORA MPJPE ↓	AGORA MVE ↓
video	HMMR [19]	116.5	72.6	-	-	56.9	-	-
	TCMR [7]	86.5	52.7	102.9	-	-	-	-
	VIBE [22]	82.7	51.9	99.1	65.6	41.1	-	-
	MAED [58]	79.1	45.7	92.6	56.4	38.7	-	-
model-free	I2L-MeshNet [36]	93.2	58.6	110.1	-	-	-	-
	Pose2Mesh [8]	89.5	56.3	105.3	64.9	46.3	-	-
	HybrIK [27]	80.0	48.8	94.5	54.4	34.5	-	-
	METRO [28]	77.1	47.9	88.2	54.0	36.7	-	-
	Graphormer [29]	74.7	45.6	87.7	51.2	34.5	-	-
model-based	HMR [18]	130.0	81.3	-	-	56.8	180.5	173.6
	SPIN [25]	96.9	59.2	116.4	-	41.1	153.4	148.9
	SPEC [24]	96.5	53.2	118.5	-	-	112.3	106.5
	HMR-EFT [17]	85.1	52.2	98.7	63.2	43.8	165.4	159.0
	PARE [23]	79.1	46.4	94.2	-	-	146.2	140.9
	ROMP [54]	76.7	47.3	93.4	-	-	116.6	113.8
	CLIFF (Res-50)	72.0 -13.1	45.7	85.3	50.5	35.1	91.7	86.3
	CLIFF (HR-W48)	**69.0** -5.7	**43.0**	**81.2**	**47.1**	**32.7**	**81.0**	**76.0**

- CLIFF gets better pixel-alignments in full images



5. Conclusion

Disclosure
global rotations cannot be accurately inferred when only using cropped images

Model
CLIFF, a model fed and supervised with global-location-aware information

Annotator
CLIFF+, a novel pseudo-ground-truth annotator

Reference:

Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)