End-to-end Recovery of Human Shape and Pose

July 26, 2023

Presenter, Soyeong Sohn

Abstract

- End-to-end framework for reconstructing a full 3D mesh of a human body from a single RGB image
- Produce a richer and more useful mesh representation that is parameterized by shape and 3D joint angles
- The main objective: minimize the reprojection loss of keypoints -> able to be trained with datasets only have 2D annotations
- Infer 3D pose and shape parameters directly from image pixels

Introduction

Problems on existing methods

- Most approaches focus on recovering 3D joint locations, but joints alone are not the full story
- Joints alone do not constrain the full DoF at each joint
- Existing methods for recovering 3D human mesh focus on a multi-stage approach, which is not optimal
    - Estimate 2D joint location -> estimate the 3D model parameters

-> Propose an end-to-end solution to learn a mapping from image pixels directly to model parameters

Introduction

Challenges in training a model in an end-to-end manner

1. Lack of large scale ground truth 3D annotation for in-the-wild images
2. Inherent ambiguities in single-view 2D-to-3D mapping (i.e. depth ambiguity)

-> Use large-scale 2D keypoint annotations of in-the-wild images and a separate large-scale dataset of 3D meshes of people with various pose and shapes

Contribution

- The first end-to-end framework for reconstructing 3D mesh of human body from a single RGB image
- Train model with the datasets which only have ground truth 2D annotations by the reprojection loss
- Always infer the full 3D body even in cases of occlusion and truncation
- Take advantage of unpaired 2D keypoint annotations and 3D scans in a conditional generative adversarial manner

Model - Overview

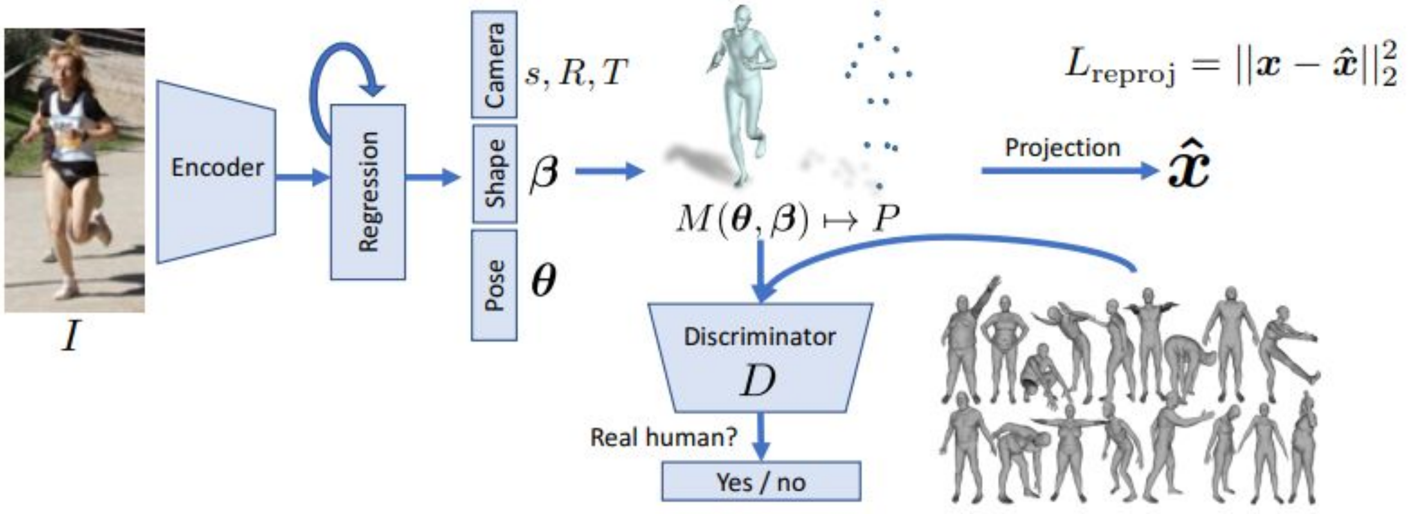


Figure 2: **Overview of the proposed framework.** An image I is passed through a convolutional encoder. This is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters are also sent to the discriminator D, whose goal is to tell if these parameters come from a real human shape and pose.

Model

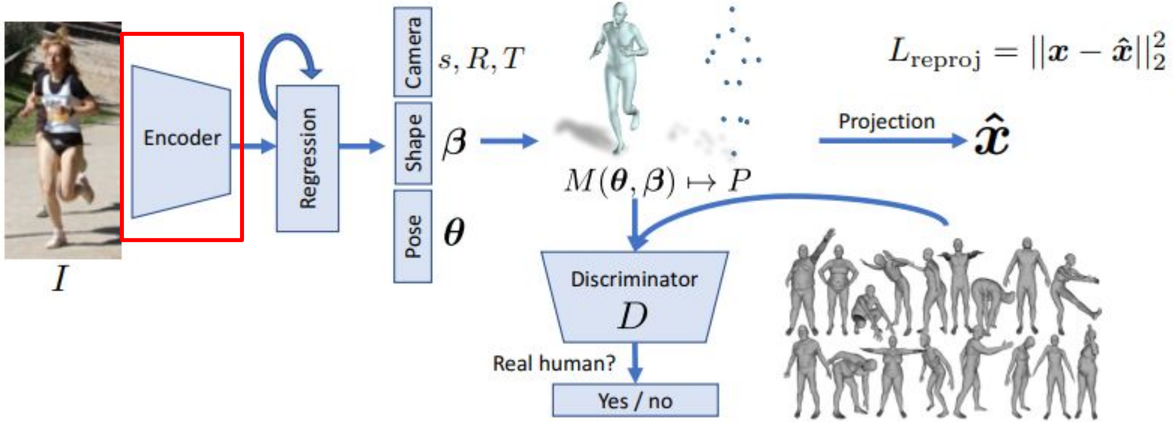


Figure 2: **Overview of the proposed framework.** An image I is passed through a convolutional encoder. This is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters are also sent to the discriminator D, whose goal is to tell if these parameters come from a real human shape and pose.

- CNN encoder: ResNet-50 pretrained on ImageNet classification task

Model

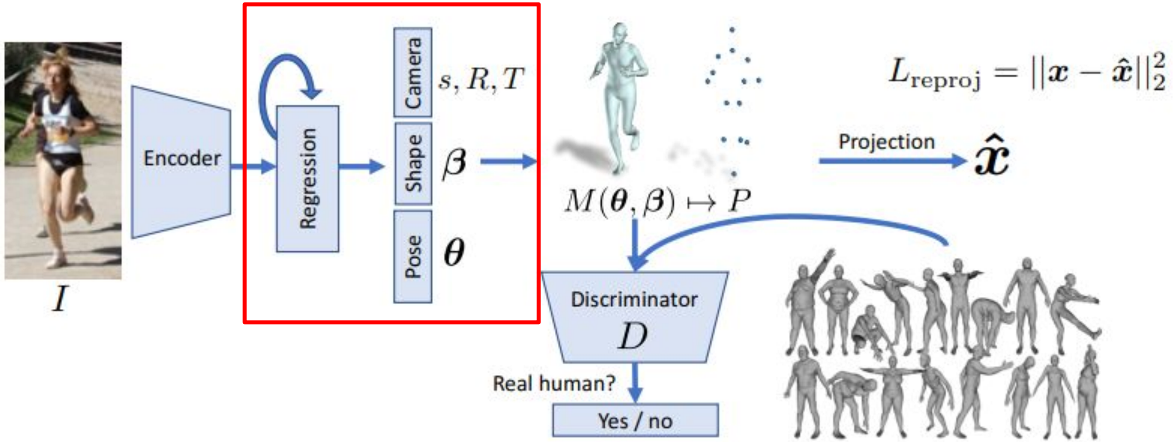


Figure 2: **Overview of the proposed framework.** An image I is passed through a convolutional encoder. This is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters are also sent to the discriminator D, whose goal is to tell if these parameters come from a real human shape and pose.

- 3D reconstruction of human body: 85-d vector
- θ: pose (3 x 23(# joints)=69-d)
- β: shape (10-d; first 10 coef of a PCA shape space)

- R: global rotation in axis-angle representation (3-d)
- t: translation (2-d)
- s: scale (1-d)

Model

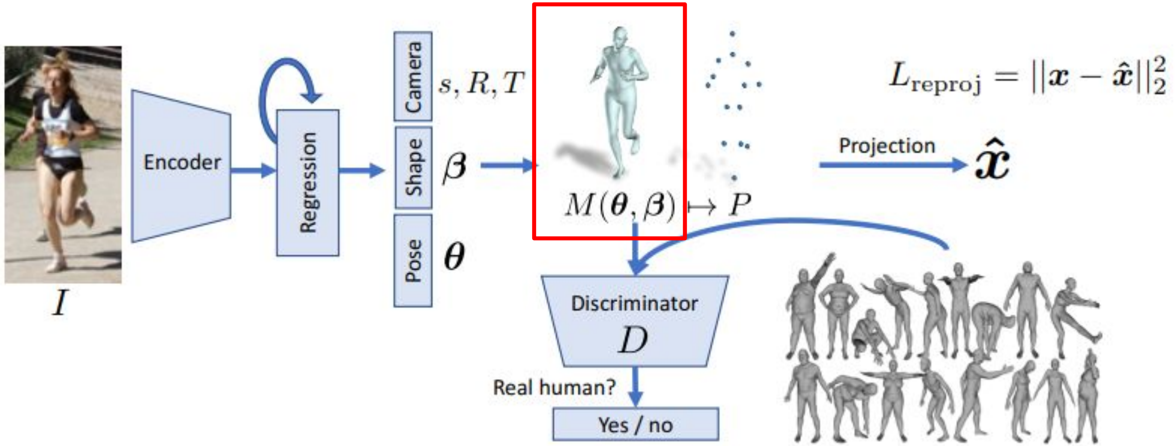
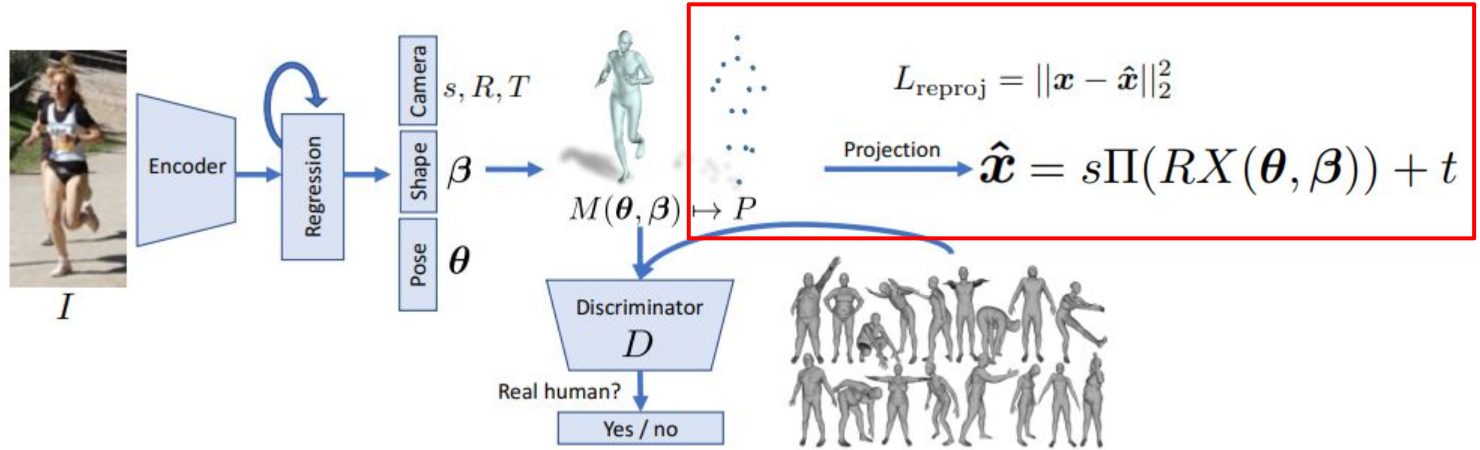


Figure 2: **Overview of the proposed framework.** An image I is passed through a convolutional encoder. This is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters are also sent to the discriminator D, whose goal is to tell if these parameters come from a real human shape and pose.

- $M(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{3 \times N}$: a triangulated mesh with N=6980 vertices
- obtained by SMPL conditioned on θ and β

Model

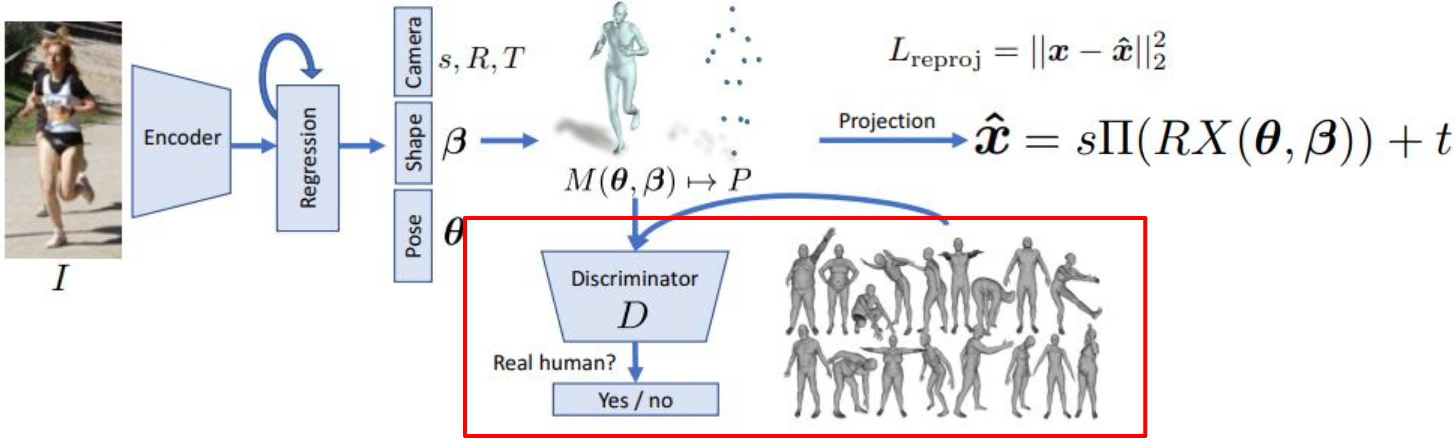


I
Encoder
Regression
Camera
Shape
Pose
s, R, T
$\boldsymbol{\beta}$
$\boldsymbol{\theta}$
$M(\boldsymbol{\theta}, \boldsymbol{\beta}) \mapsto P$
Discriminator
D
Real human?
Yes / no
Projection
$L_{\text{reproj}} = ||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2^2$
$\hat{\boldsymbol{x}} = s\Pi(RX(\boldsymbol{\theta}, \boldsymbol{\beta})) + t$

Figure 2: **Overview of the proposed framework.** An image I is passed through a convolutional encoder. This is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters are also sent to the discriminator D, whose goal is to tell if these parameters come from a real human shape and pose.

- $X(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{3\times P}$: 3D keypoints used for reprojection error; obtained by linear regression from M(θ, β)
- Reprojecting 3D keypoints to 2D space(orthographic projection) so that we can compute loss with 2D g.t. keypoints

Model



I
Encoder
Regression
Camera
Shape
Pose
s, R, T
β
θ
$M(\theta, \beta) \mapsto P$
$L_{\text{reproj}} = ||x - \hat{x}||_2^2$
Projection
$\hat{x} = s\Pi(RX(\theta, \beta)) + t$
Discriminator
D
Real human?
Yes / no

Figure 2: **Overview of the proposed framework.** An image I is passed through a convolutional encoder. This is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters are also sent to the discriminator D, whose goal is to tell if these parameters come from a real human shape and pose.

- Anthropometrically implausible 3D bodies or bodies with gross self-intersection may still minimize the reprojection loss
- Use a discriminator network to regularize this

Model - Loss Functions

Overall Objective

$$L = \lambda(L_{\text{reproj}} + \mathbb{1} L_{\text{3D}}) + L_{\text{adv}}$$

- λ: relative importance of each objective
- $\mathbb{1}$: whether g.t 3D is available or not

$$L_{\text{reproj}} = \Sigma_i ||v_i(\mathbf{x}_i - \hat{\mathbf{x}}_i)||_1,$$

- L1 loss between g.t. 2D joints and estimated 2D joints
- v_i: visibility for each joints (1 if visible else 0)

$$L_{\text{3D}} = L_{\text{3D joints}} + L_{\text{3D smpl}}$$

$$L_{\text{joints}} = ||(\mathbf{X_i} - \hat{\mathbf{X}}_\mathbf{i})||_2^2$$

$$L_{\text{smpl}} = ||[\boldsymbol{\beta}_i, \boldsymbol{\theta}_i] - [\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\theta}}_i]||_2^2.$$

- Only used when 3D supervision is employed

$$\min L_{\text{adv}}(E) = \sum_i \mathbb{E}_{\Theta \sim p_E}[(D_i(E(I)) - 1)^2]$$

$$\min L(D_i) = \mathbb{E}_{\Theta \sim p_{\text{data}}}[(D_i(\Theta) - 1)^2] + \mathbb{E}_{\Theta \sim p_E}[D_i(E(I))^2]$$

Experiments

- Evaluate quantitatively on the standard 3D joint estimation task since there are no ground truth mesh 3D annotations for current datasets
- Qualitative comparison on images from MS COCO with occlusion, cluster, truncation, and complex poses
- Quantitative comparison on Human3.6M, MPI-INF-3DHP, LSP and MS COCO at various error percentiles

Experiments

Qualitative Results

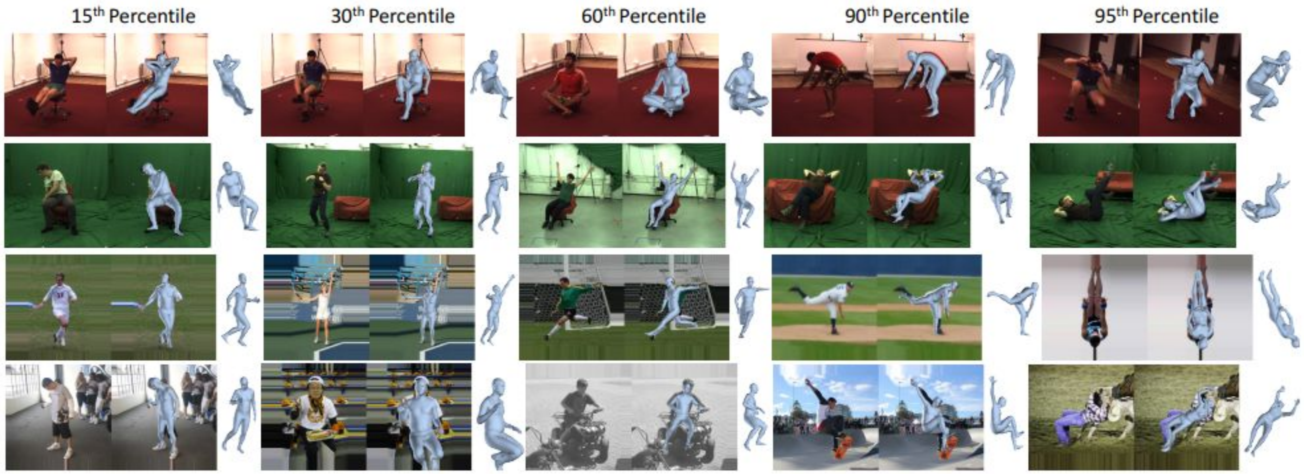
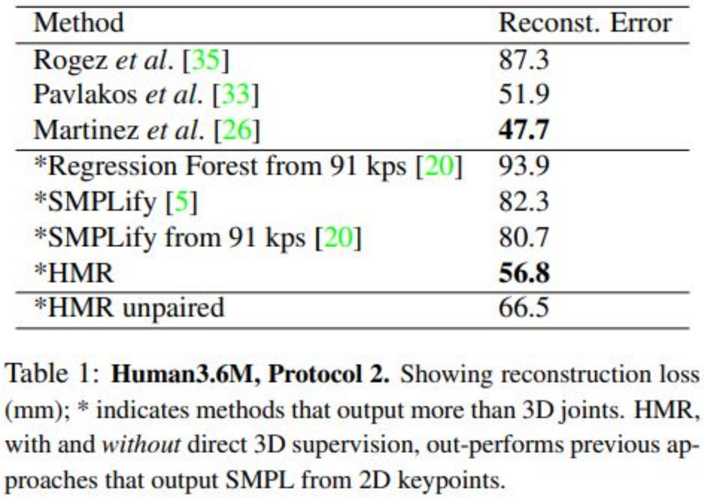


Figure 3: **Results sampled from different datasets at the 15th, 30th, 60th, 90th and 95th error percentiles.** Percentiles are computed using MPJPE for 3D datasets (first two rows - Human3.6M and MPI-INF-3DHP) and 2D pose PCK for 2D datasets (last two rows - LSP and MS COCO). High percentile indicates high error. Note results at high error percentile are often semantically quite reasonable.

- MPJPE does not appear to correlate well with the visual quality of the results

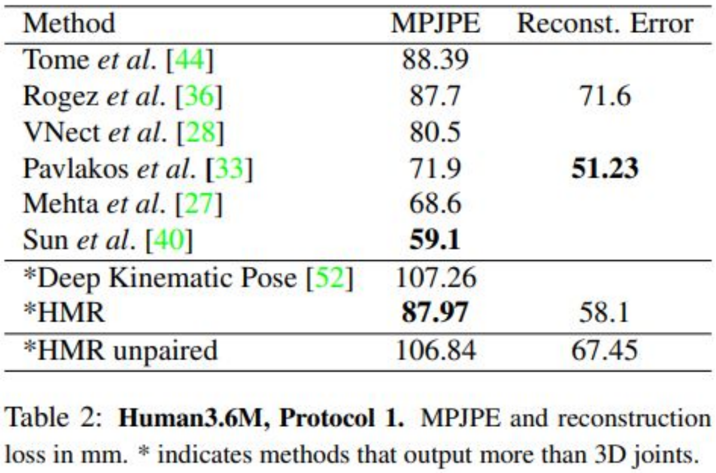
Experiments

Quantitative Results



Method	Reconst. Error
Rogez *et al.* [35]	87.3
Pavlakos *et al.* [33]	51.9
Martinez *et al.* [26]	**47.7**
*Regression Forest from 91 kps [20]	93.9
*SMPLify [5]	82.3
*SMPLify from 91 kps [20]	80.7
*HMR	**56.8**
*HMR unpaired	66.5

Table 1: **Human3.6M, Protocol 2.** Showing reconstruction loss (mm); * indicates methods that output more than 3D joints. HMR, with and *without* direct 3D supervision, out-performs previous approaches that output SMPL from 2D keypoints.



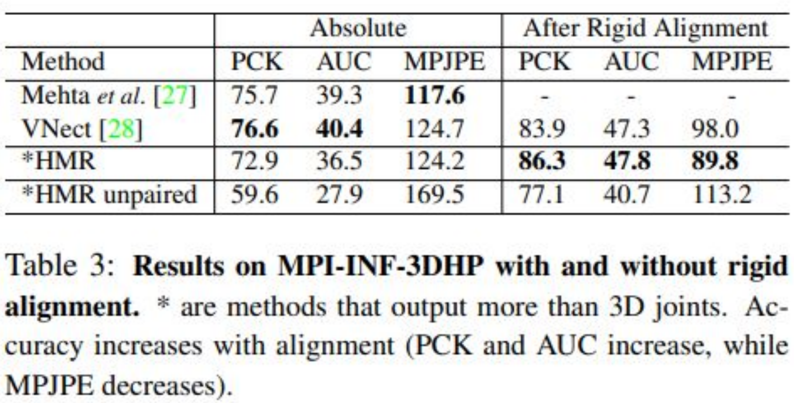
Method	MPJPE	Reconst. Error
Tome *et al.* [44]	88.39	
Rogez *et al.* [36]	87.7	71.6
VNect *et al.* [28]	80.5	
Pavlakos *et al.* [33]	71.9	**51.23**
Mehta *et al.* [27]	68.6	
Sun *et al.* [40]	**59.1**	
*Deep Kinematic Pose [52]	107.26	
*HMR	**87.97**	58.1
*HMR unpaired	106.84	67.45

Table 2: **Human3.6M, Protocol 1.** MPJPE and reconstruction loss in mm. * indicates methods that output more than 3D joints.

- Protocol 1: trained on 5 subjects(S1, S5, S6, S7, S8) and tested on 2 (S9, S11)
- Protocol 2: same train/test set, tested only on the frontal camera and reports reconstruction error

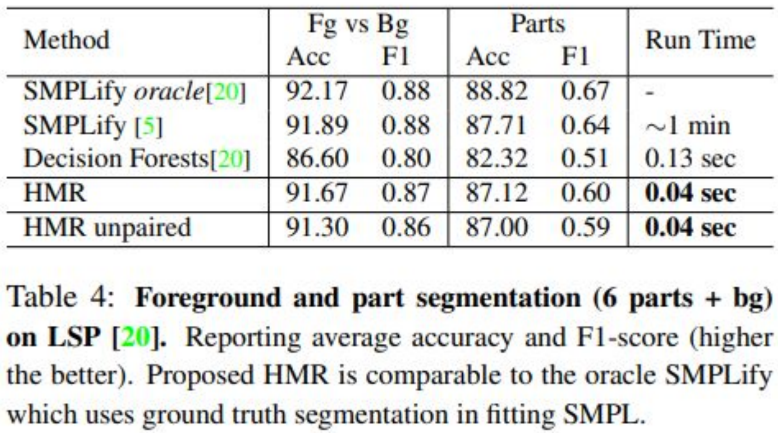
Experiments

Quantitative Results



	Absolute			After Rigid Alignment		
Method	PCK	AUC	MPJPE	PCK	AUC	MPJPE
Mehta *et al.* [27]	75.7	39.3	**117.6**	-	-	-
VNect [28]	**76.6**	**40.4**	124.7	83.9	47.3	98.0
*HMR	72.9	36.5	124.2	**86.3**	**47.8**	**89.8**
*HMR unpaired	59.6	27.9	169.5	77.1	40.7	113.2

Table 3: **Results on MPI-INF-3DHP with and without rigid alignment.** * are methods that output more than 3D joints. Accuracy increases with alignment (PCK and AUC increase, while MPJPE decreases).



Method	Fg vs Bg		Parts		Run Time
	Acc	F1	Acc	F1	
SMPLify *oracle*[20]	92.17	0.88	88.82	0.67	-
SMPLify [5]	91.89	0.88	87.71	0.64	~1 min
Decision Forests[20]	86.60	0.80	82.32	0.51	0.13 sec
HMR	91.67	0.87	87.12	0.60	**0.04 sec**
HMR unpaired	91.30	0.86	87.00	0.59	**0.04 sec**

Table 4: **Foreground and part segmentation (6 parts + bg) on LSP [20].** Reporting average accuracy and F1-score (higher the better). Proposed HMR is comparable to the oracle SMPLify which uses ground truth segmentation in fitting SMPL.

Experiments

Other Experiments Results

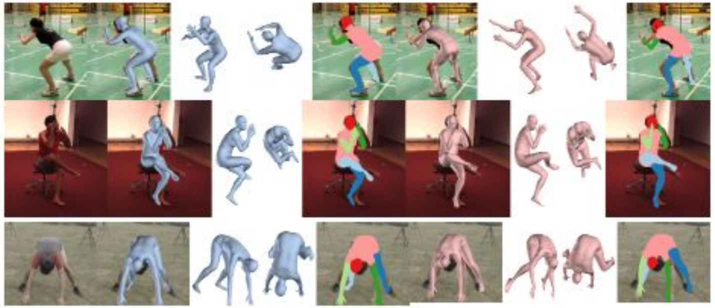


Figure 4: **Results with and without paired 3D supervision.** 3D reconstructions, without direct 3D supervision, are very close to those of the supervised model.

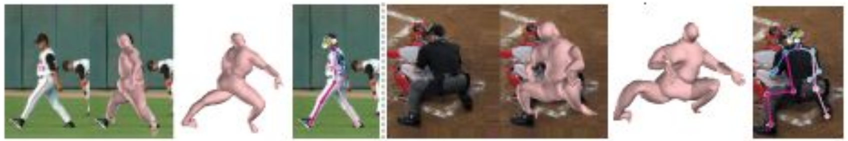


Figure 5: **No Discriminator No 3D.** With neither the discriminator, nor the direct 3D supervision, the network produces monsters. On the right of each example we visualize the ground truth keypoint annotation in unfilled circles, and the projection in filled circles. Note that despite the unnatural pose and shape, its 2D projection error is very accurate.

UNIST

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY
2007

THANK YOU