VIBE: Video Inference for Human Body Pose and Shape Estimation

Muhammed Kocabas[1,2], Nikos Athanasiou[1], Michael J. Black[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany

[2] Max Planck ETH Center for Learning Systems

CVPR 2020

Guray Ozgur

University of Tübingen

Agenda

- Introduction
  - Goal and Motivation
  - Prior Work
- **VIBE: V**ideo **I**nference for Human **B**ody Pose and Shape **E**stimation
  - Method overview
  - Implementation details
  - Ablation study
- Results
  - Qualitative Results
  - Quantitative Results
  - Limitations
- Summary

Introduction

- Goal
- Motivation
- Prior Work

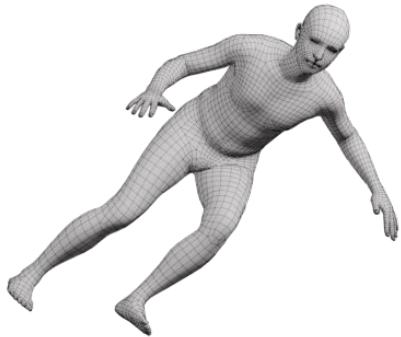
Goal

- **Estimating 3D** pose and shape of a person given a **video**
- Creating 3D avatars that **know how to move**
- Producing **accurate** and **natural** motion sequences

Kocabas et al. CVPR'20.

Motivation

- **Movement involved** in **interactions** with the world and with each other
- **Understanding** human behaviour thru human motions
- **Serving as** additional **training data** for motion synthesis



Kocabas et al. CVPR'20.

Prior Work on Motion Estimation

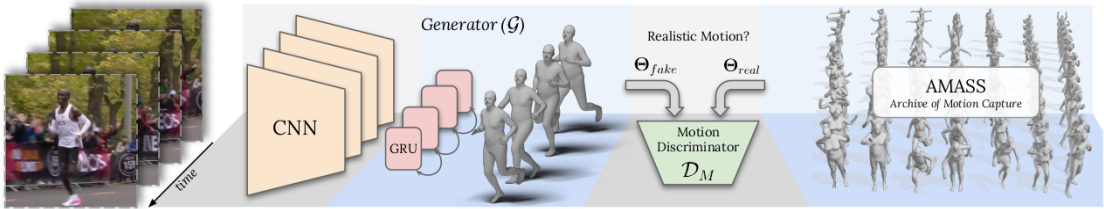
- 3D pose and shape from a single image
  - Key-point detections to SMPL body model
  - Pixels to SMPL body model
  - **Problem:** Jittery, unstable results
- 3D pose and shape from video
  - ”Lifting” 2D key-points to 3D
  - Temporal models to use more information from video
- GANs for sequence modeling
  - To predict future motion sequences
  - To generate new motion sequences
- Closest work

  T-HMR (Temporal Human Mesh Recovery), Kanazawa et al. 18’: End-to-end Recovery of Human Shape and Pose

Kocabas et al. CVPR’20.

VIBE: Video **I**nference for Human **B**ody Pose and Shape **E**stimation

- Method overview
- Implementation details
- Ablation study

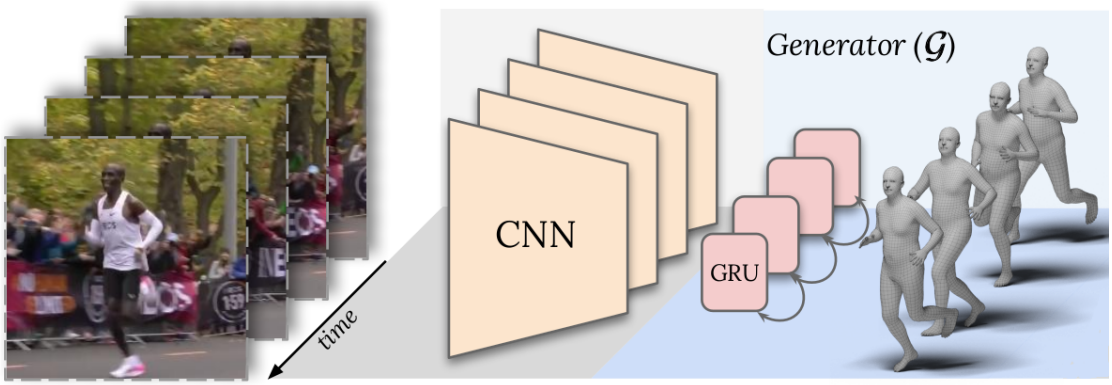


Kocabas et al. CVPR'20.

What is it?

- Given
  - challenging in-the-wild videos
- Extracting features of each frame using a pretrained CNN
- Temporal encoder using a recurrent architecture
- Regression of encodings to SMPL body model
- Adversarial training with a real dataset of human motions

Generator



Kocabas et al. CVPR'20.

Temporal Encoder

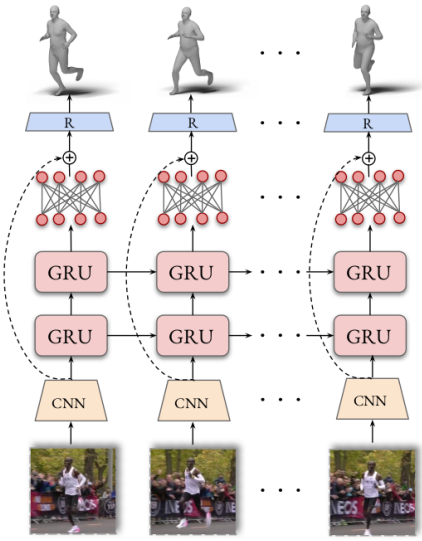
$$L_{\mathcal{G}} = L_{3D} + L_{2D} + L_{SMPL} + L_{adv}$$

$$L_{3D} = \sum_{t=1}^{T} \|X_t - \hat{X}_t\|_2,$$

$$L_{2D} = \sum_{t=1}^{T} \|x_t - \hat{x}_t\|_2,$$

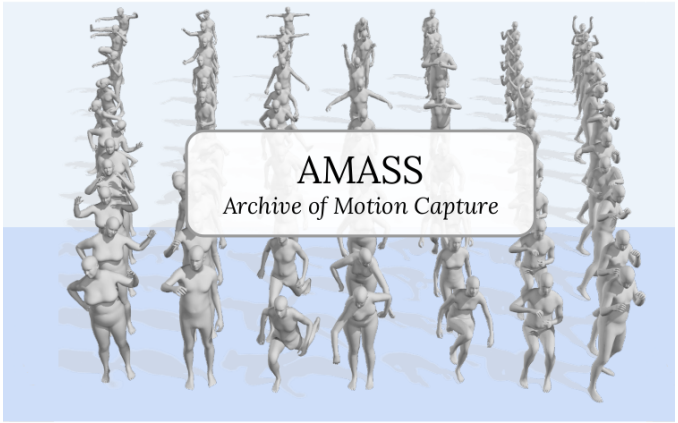
$$L_{SMPL} = \|\beta - \hat{\beta}\|_2 + \sum_{t=1}^{T} \|\theta_t - \hat{\theta}_t\|_2$$

$$L_{adv} = \mathbb{E}_{\Theta \sim p_G}[(\mathcal{D}_M(\hat{\mathbf{\Theta}}) - 1)^2]$$



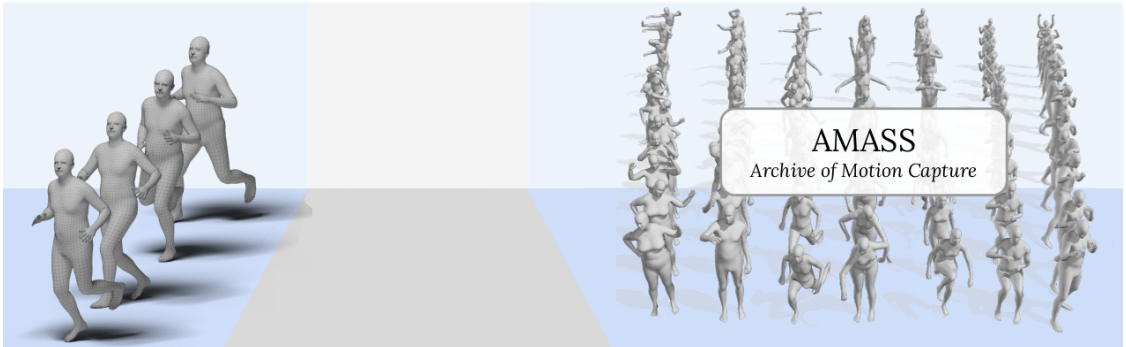
Kocabas et al. CVPR'20.

AMASS



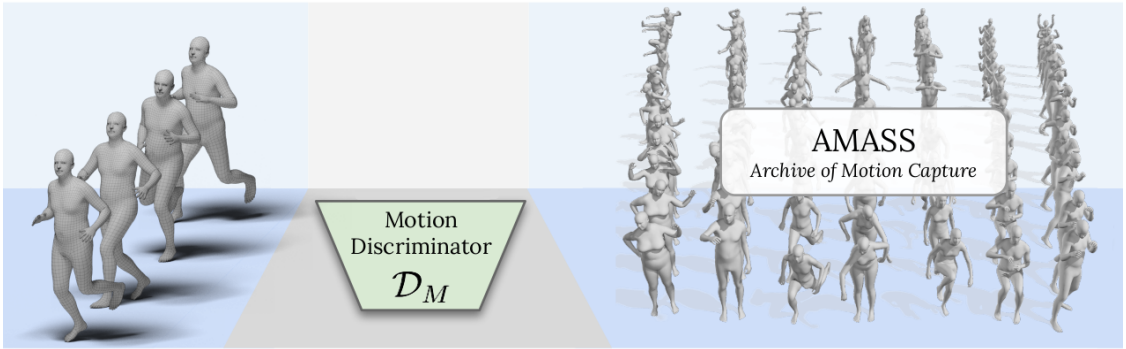
Kocabas et al. CVPR'20.

Discriminator



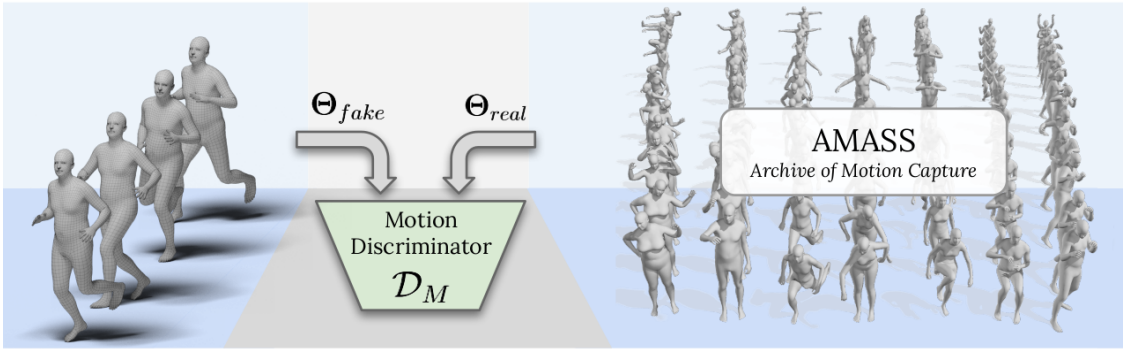
Kocabas et al. CVPR'20.

Discriminator



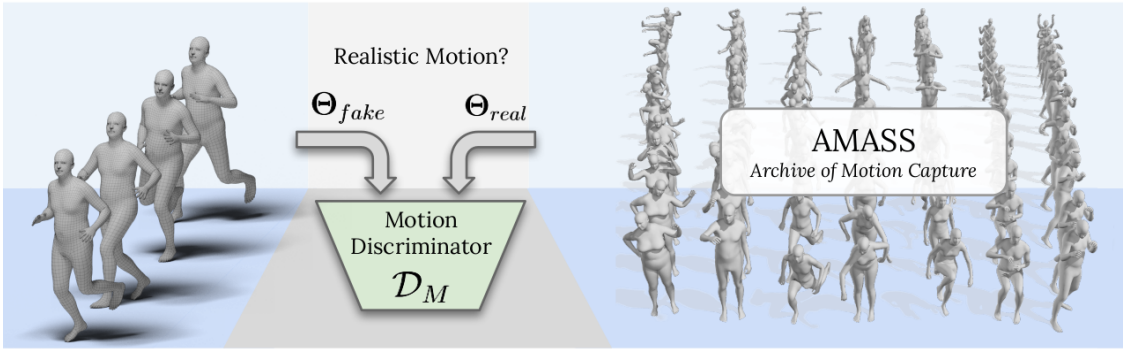
Kocabas et al. CVPR'20.

Discriminator



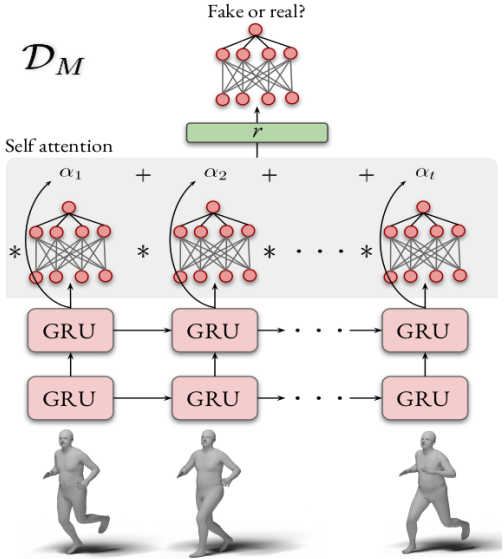
Kocabas et al. CVPR'20.

Discriminator



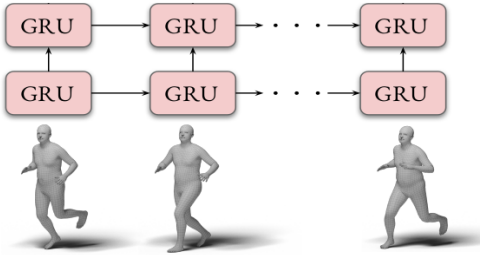
Kocabas et al. CVPR'20.

Motion discriminator architecture



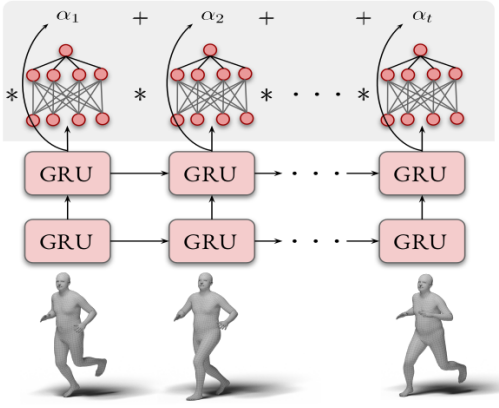
Kocabas et al. CVPR'20.

GRU layers



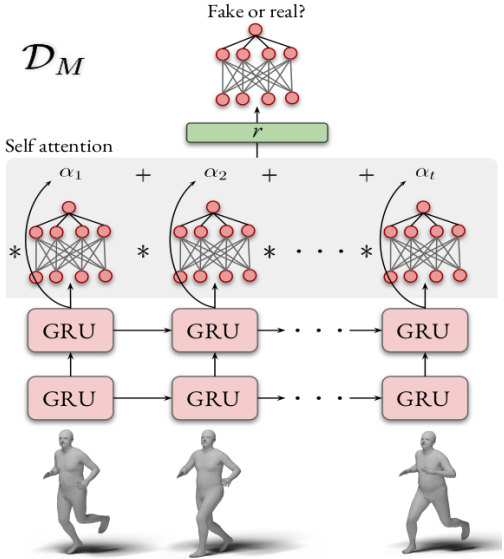
Kocabas et al. CVPR'20.

Self attention layer



Kocabas et al. CVPR'20.

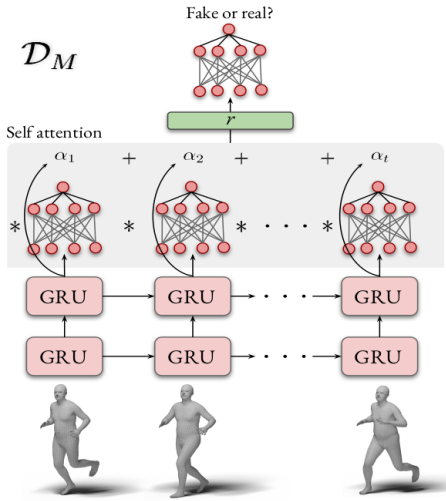
Real/Fake probability for each input



Kocabas et al. CVPR'20.

Motion discriminator architecture

$$L_{\mathcal{D}_M} = \mathbb{E}_{\Theta \sim p_R}[(\mathcal{D}_M(\mathbf{\Theta}) - 1)^2] + \mathbb{E}_{\Theta \sim p_G}[\mathcal{D}_M(\hat{\mathbf{\Theta}})^2]$$



Training Data: Batches of mixed 2D and 3D datasets

- **Ground-truth 2D video datasets:** PennAction and PoseTrack
- **Pseudo ground-truth datasets annotated using a 2D key-point detector:** InstaVariety and Kinetics-400
- **3D joint labels:** MPI-INF-3DHP and Human3.6M
- **Adversarial training:** AMASS

Implementation details

- **Pose Generator**
  - feature extraction using ResNet50
  - a 2-layer GRU with a hidden size of 1024 followed by a linear projection layer
  - a SMPL parameter regressor initilized with pre-trained weights from HMR
  - residual connections
- **Motion Discriminator**
  - a 2-layer GRU with a hidden size of 1024 and tanh activation
  - a self-attention mechanism with 2 MLP layers, each with 1024 neurons, and a dropout rate of 0.1
- Adam optimizer with a learning rate of $5x10^{-5}$ and $1x10^{-4}$ for generator and discriminator resp.

Ablation study

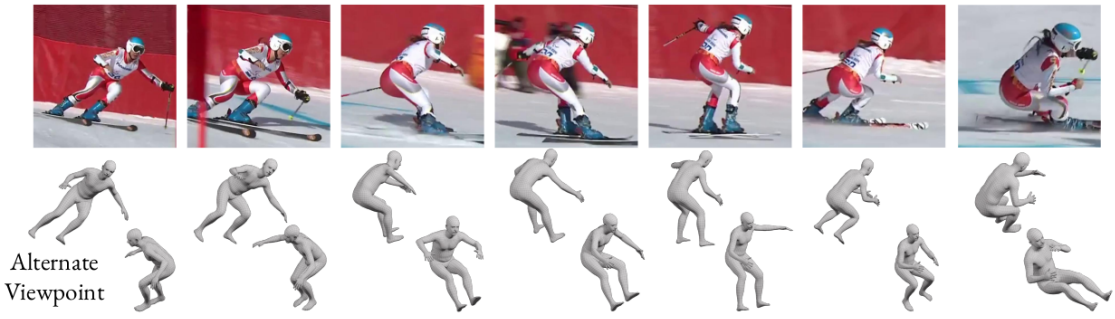
	3DPW			
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓
Kanazawa *et al*. [30]	73.6	120.1	142.7	34.3
Baseline (only $\mathcal{G}$)	75.8	126.1	147.5	28.3
$\mathcal{G}$ + $\mathcal{D}_M$	**72.4**	**116.7**	**132.4**	**27.8**
Kolotouros *et al*. [37]	60.1	102.4	129.2	29.2
Baseline (only $\mathcal{G}$)	56.9	90.2	109.5	28.0
$\mathcal{G}$ + MPoser Prior	54.1	87.0	103.9	28.2
$\mathcal{G}$ + $\mathcal{D}_M$ (VIBE)	**51.9**	**82.9**	**99.1**	**23.4**

Model	PA-MPJPE ↓	MPJPE ↓
$\mathcal{D}_M$ - concat	53.7	85. 9
$\mathcal{D}_M$ - attention [2 layers,512 nodes]	54.2	86.6
$\mathcal{D}_M$ - attention [2 layers,1024 nodes]	**51.9**	**82.9**
$\mathcal{D}_M$ - attention [3 layers,512 nodes]	53.6	85.3
$\mathcal{D}_M$ - attention [3 layers,1024 nodes]	52.4	82.7

Results

- Qualitative Results
- Quantitative Results
- Limitations

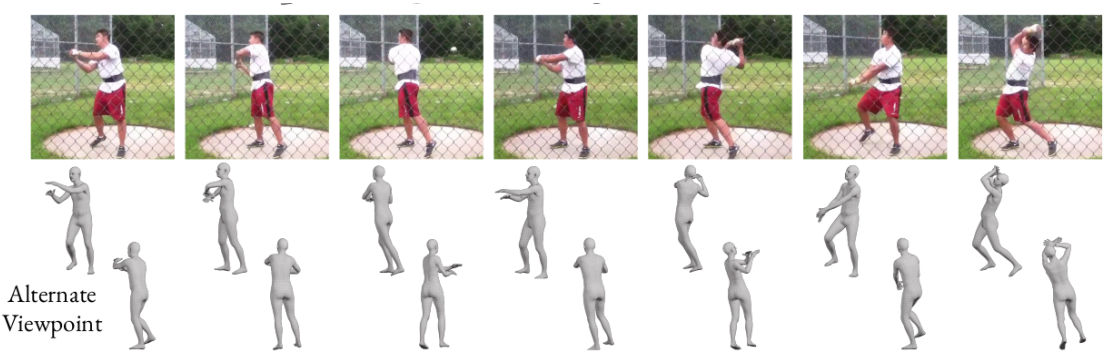
Qualitative Results



Kocabas et al. CVPR'20.

Qualitative Results

Qualitative Results



Kocabas et al. CVPR'20.

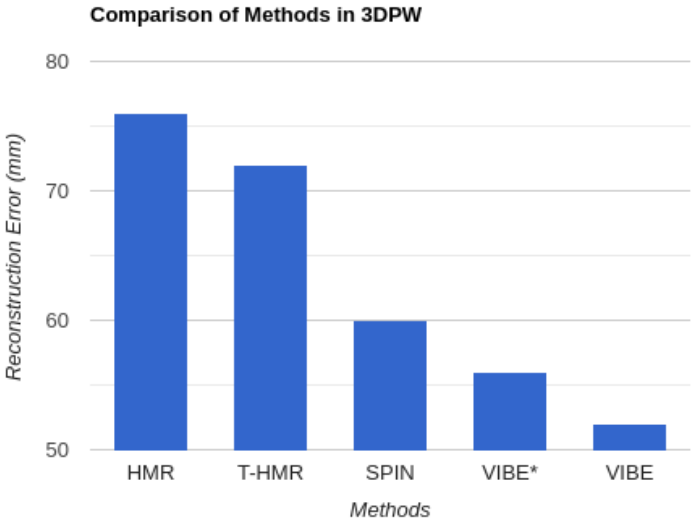
Metrics

- **MPJPE (Mean Per Joint Position Error):** distance between the ground-truth and the predicted joint positions
  - after centering the pelvis joint on the ground truth location
- **PA-MPJPE (Pro- crustes Aligned MPJPE):**
  - after a rigid alignment of the predicted pose to the end ground-truth pose
- **PVE (Per-Vertex-Error):** distance between the ground-truth and predicted mesh vertices
- **PCK (Percentage of Correct Keypoints):** ratio of correct cases where the distance between the actual and predicted joint positions is below a predefined threshold
- **Accel (Acceleration Error):** difference between ground-truth and predicted 3D acceleration for every joint

Comparison with previous work

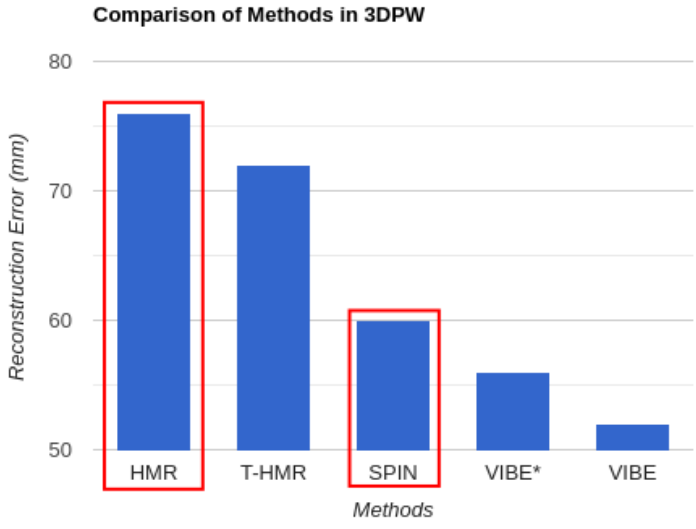
		3DPW				MPI-INF-3DHP			H36M	
	Models	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	PCK ↑	PA-MPJPE ↓	MPJPE ↓
Frame-based	Kanazawa *et al.* [30]	76.7	130.0	-	37.4	89.8	124.2	72.9	56.8	88
	Omran *et al.* [47]	-	-	-	-	-	-	-	59.9	-
	Pavlakos *et al.* [50]	-	-	-	-	-	-	-	75.9	-
	Kolotouros *et al.* [38]	70.2	-	-	-	-	-	-	50.1	-
	Arnab *et al.* [6]	72.2	-	-	-	-	-	-	54.3	77.8
	Kolotouros *et al.* [37]	59.2	96.9	116.4	29.8	67.5	105.2	76.4	**41.1**	-
Temporal	Kanazawa *et al.* [31]	72.6	116.5	139.3	**15.2**	-	-	-	56.9	-
	Doersch *et al.* [16]	74.7	-	-	-	-	-	-	-	-
	Sun *et al.* [56]	69.5	-	-	-	-	-	-	42.4	**59.1**
	VIBE (direct comp.)	56.5	93.5	113.4	27.1	**63.4**	97.7	**89.0**	41.5	65.9
	VIBE	**51.9**	**82.9**	**99.1**	23.4	64.6	**96.6**	89.3	41.4	65.6

Comparison with previous work



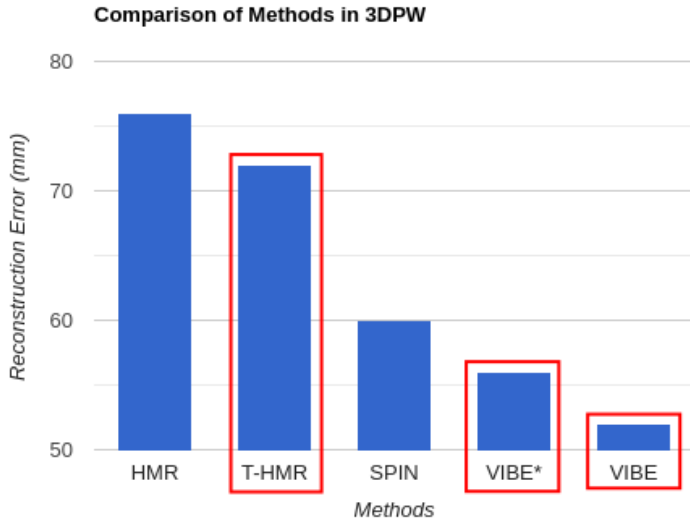
Kocabas et al. CVPR'20.

Comparison with previous work



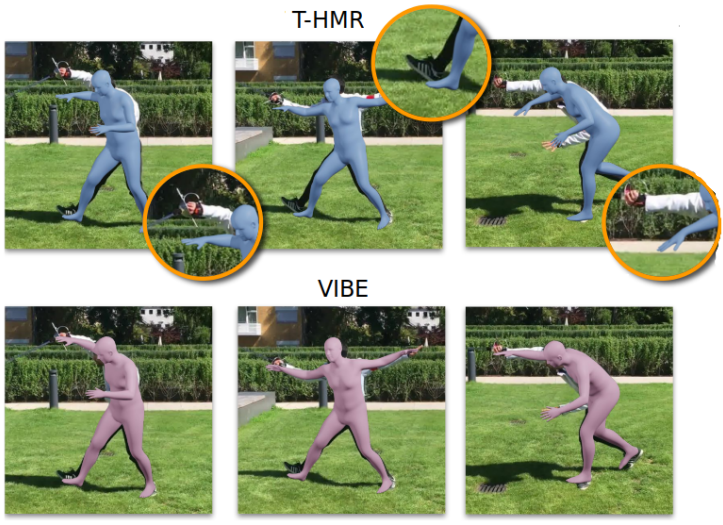
Kocabas et al. CVPR'20.

Comparison with previous work



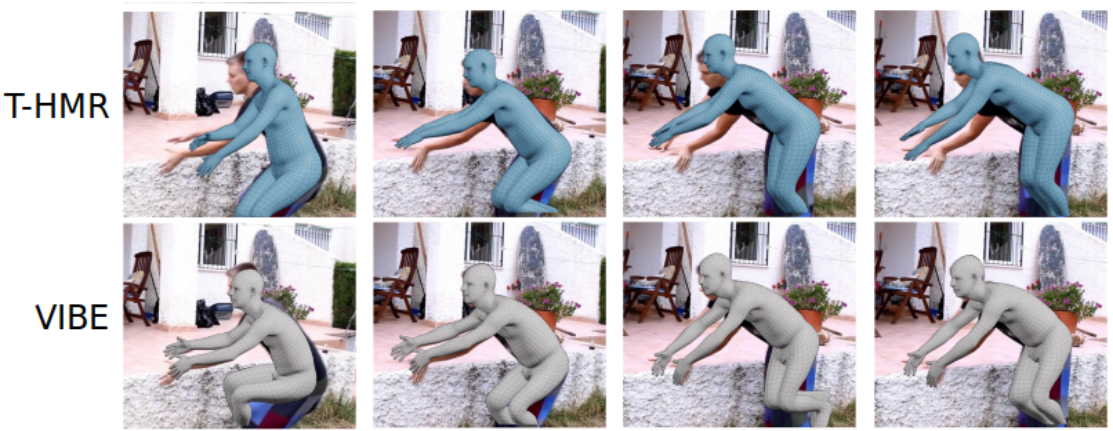
Kocabas et al. CVPR'20.

Comparison with previous work



Kocabas et al. CVPR'20.

Comparison with previous work



Kocabas et al. CVPR'20.

Limitations

- Heavy occlusions (When the person is behind something.)
- Fast actions
- Multi-person occlusions (When people are interacting with each other closely.)
- No detailed motion (Especially in hands and feet)

Kocabas et al. CVPR'20.

Summary

Summary

- a recurrent architecture that propagates information over time
- discriminative training of motion sequences using the AMASS dataset
- self-attention in the discriminator so that it learns to focus on the important temporal structure of human motion
- still a lot to do

Questions?