VIBE: Video Inference for Human Body Pose and Shape Estimation

Ji Yang

Vision and Learning Lab @ U of Alberta

Outline

- Introduction and Motivation
- Background
  - Human Mesh Recovery
- VIBE
  - Temporal Encoder
  - Motion Discriminator
  - Training Procedure
- Experiment and Results
  - Dataset
  - Experiments
  - Conclusion

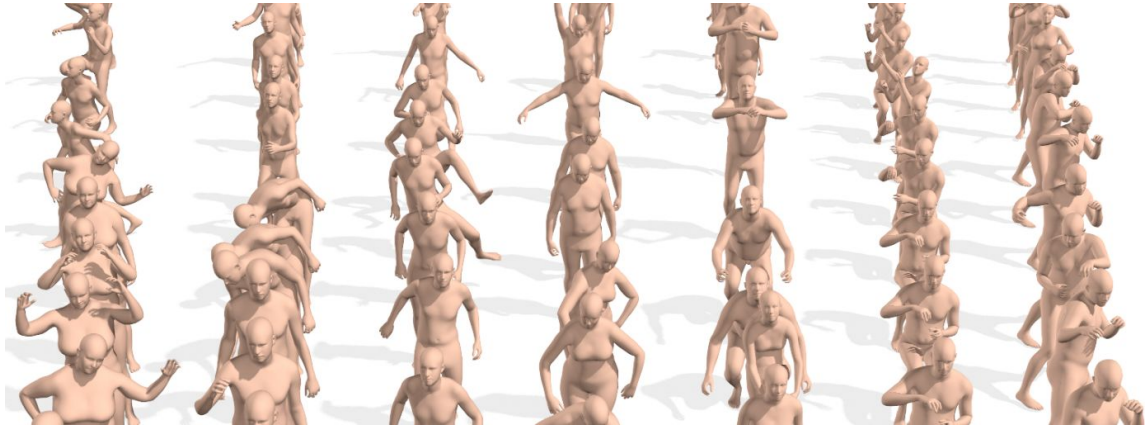
Introduction

Human motion is fundamental to understanding behavior. Despite progress on single-image 3D pose and shape estimation, existing video-based SOTA methods fail to produce accurate and natural motion sequences due to a lack of ground-truth 3D motion data for training.

The authors tries to exploit temporal information to more accurately estimate the 3D motion of the body from monocular video.

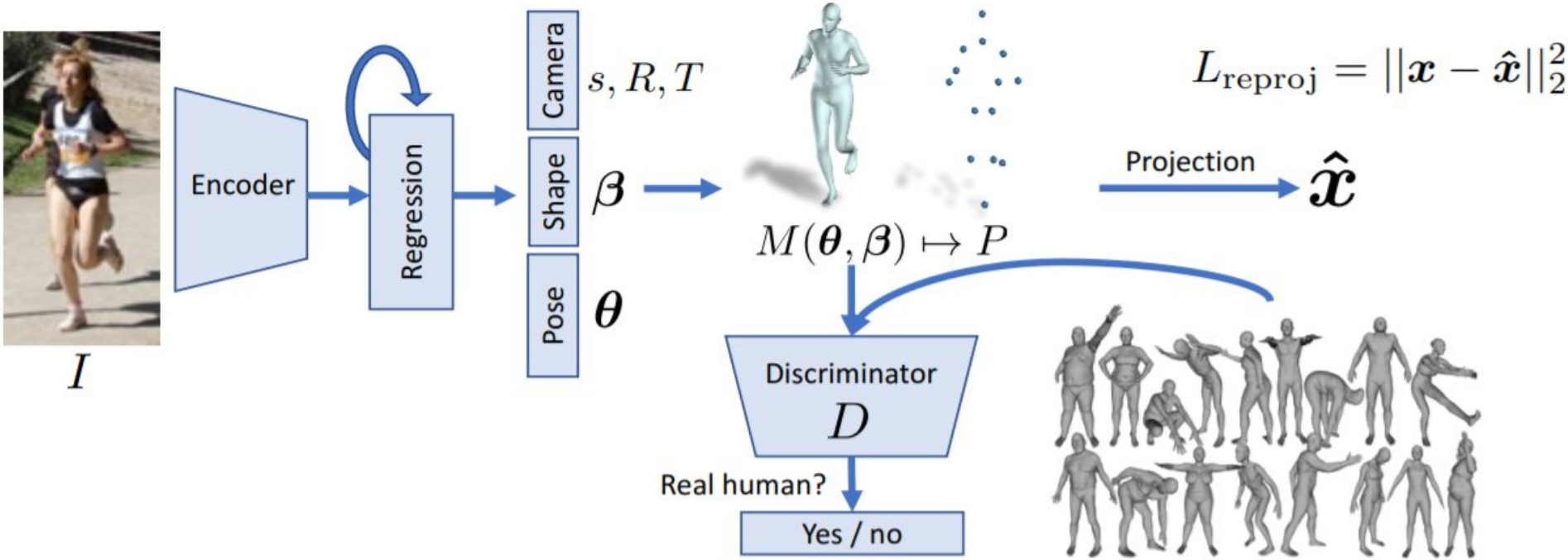
Introduction: Dataset (the primary one)

AMASS is a large database of human motion unifying different optical marker-based motion capture datasets by representing them within a common framework and parameterization. The dataset is significantly richer than previous human motion collections, having more than 40 hours of motion data, spanning over 300 subjects, more than 11000 motions. AMASS is readily useful for animation, visualization, and generating training data for deep learning.



Backgroud: HMR

Human Mesh Recovery (HMR): End-to-end adversarial learning of human pose and shape



VIBE: Overall

The human body is represented as a 3D mesh encoded using the SMPL model.

Temporal encoder (act as a Shape/Pose Generator) + Motion Discriminator

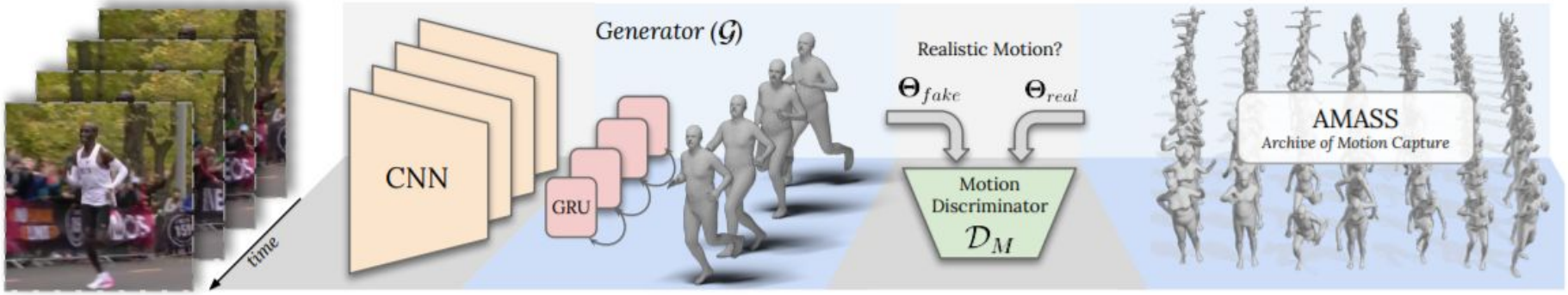


Figure 2: **VIBE architecture.** VIBE estimates SMPL body model parameters for each frame in a video sequence using a temporal generation network, which is trained together with a motion discriminator. The discriminator has access to a large corpus of human motions in SMPL format.

VIBE: Temporal Encoder

The intuition behind using a recurrent architecture is that future frames can benefit from past video information about human poses. The temporal encoder acts as a generator that, given a sequence of frames, it outputs the corresponding pose and shape of the body in each frame.

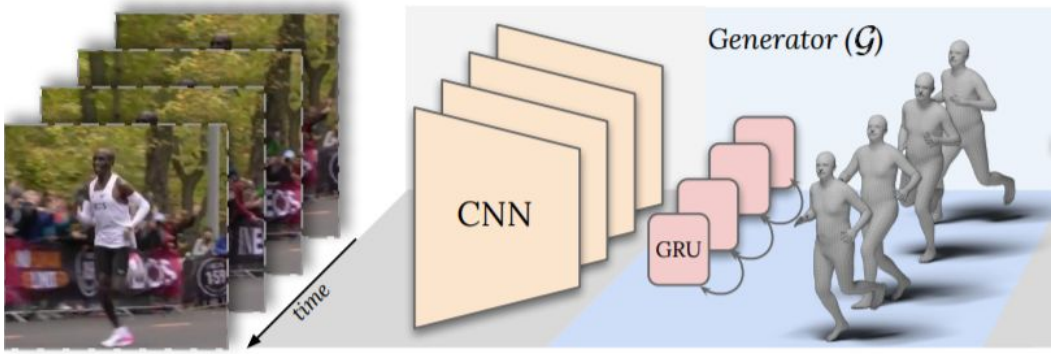
$$L_{\mathcal{G}} = L_{3D} + L_{2D} + L_{SMPL} + L_{adv} \qquad (1)$$

where each term is calculated as:

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

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

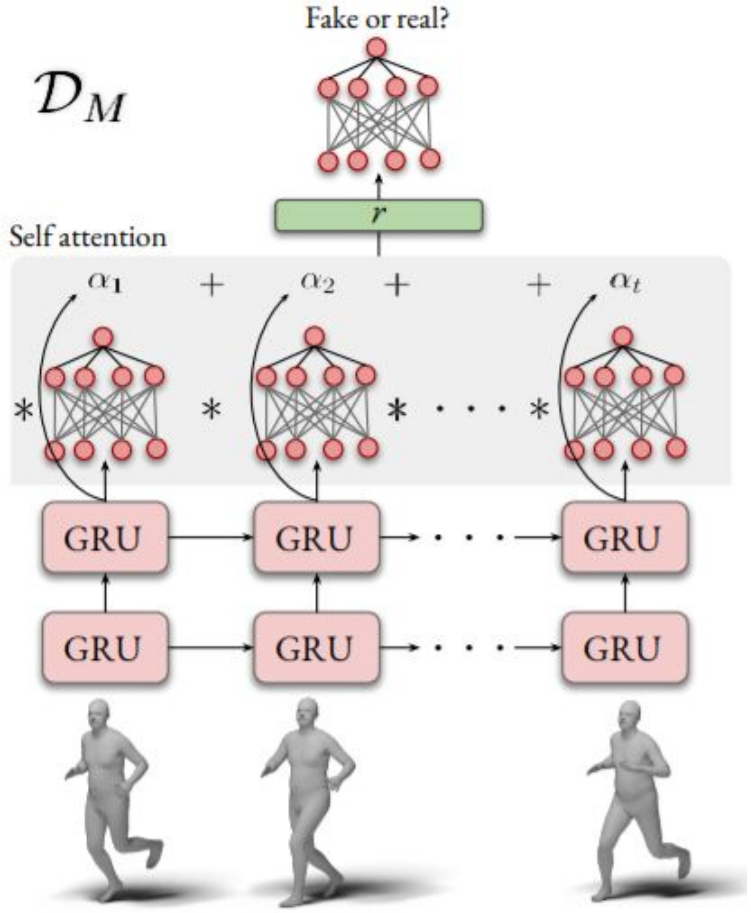
$$L_{SMPL} = \|\beta - \hat{\beta}\|_2 + \sum_{t=1}^{T} \|\theta_t - \hat{\theta}_i\|_2 \ ,$$



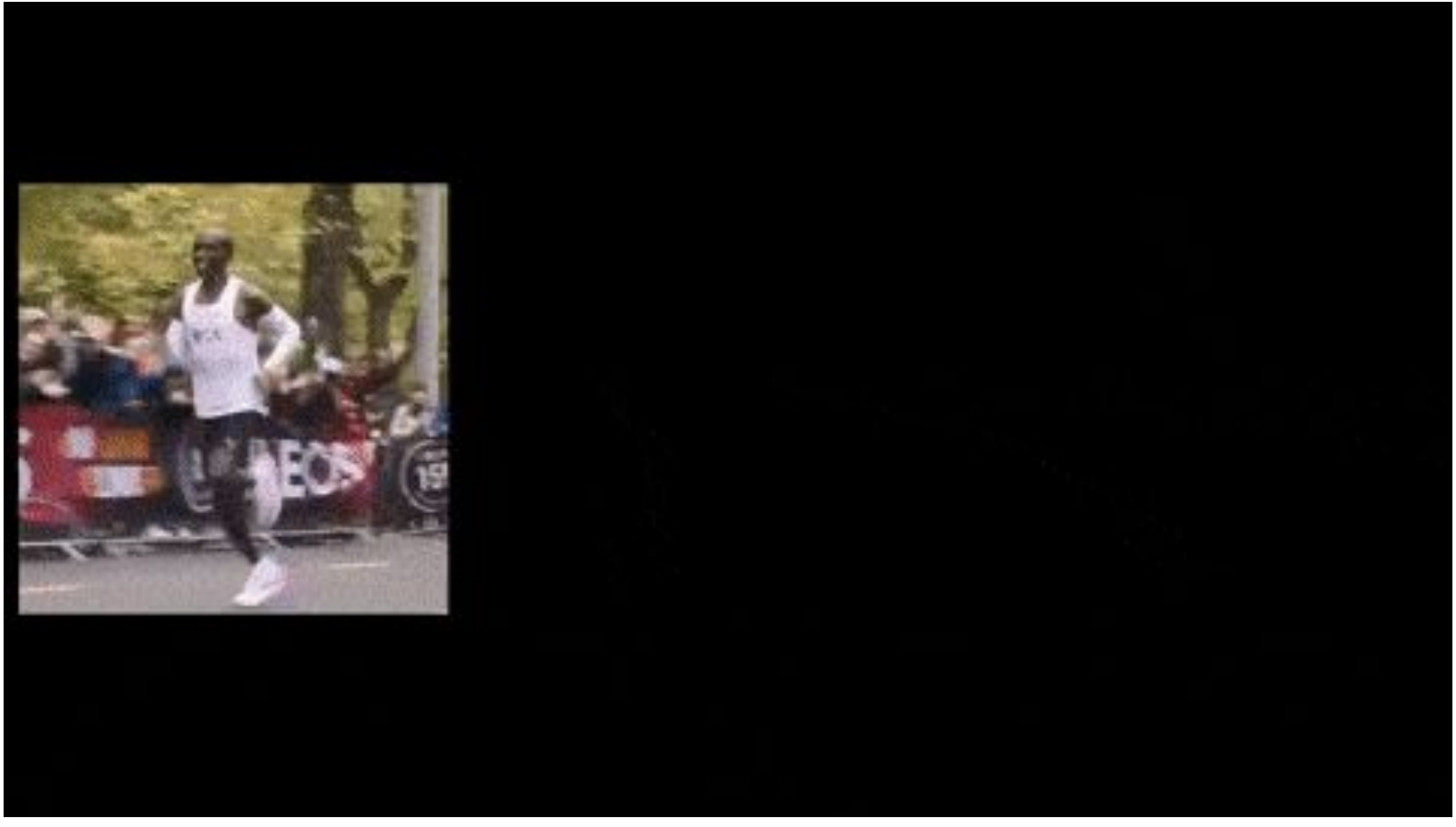
VIBE: Motion Discriminator

The output, Θ, of the generator is given as input to a multi-layer GRU model. In order to aggregate hidden states, self attention is used, followed by a linear layer that predicts a value ∈ [0, 1] representing the probability that Θ belongs to the manifold of plausible human motions.

$$L_{\mathcal{D}M} = \mathbb{E}\Theta \sim p_R \left[(\mathcal{D}M(\boldsymbol{\Theta}) - 1)^2 \right] + \mathbb{E}\Theta \sim p_G \left[\mathcal{D}_M(\hat{\boldsymbol{\Theta}})^2 \right].$$

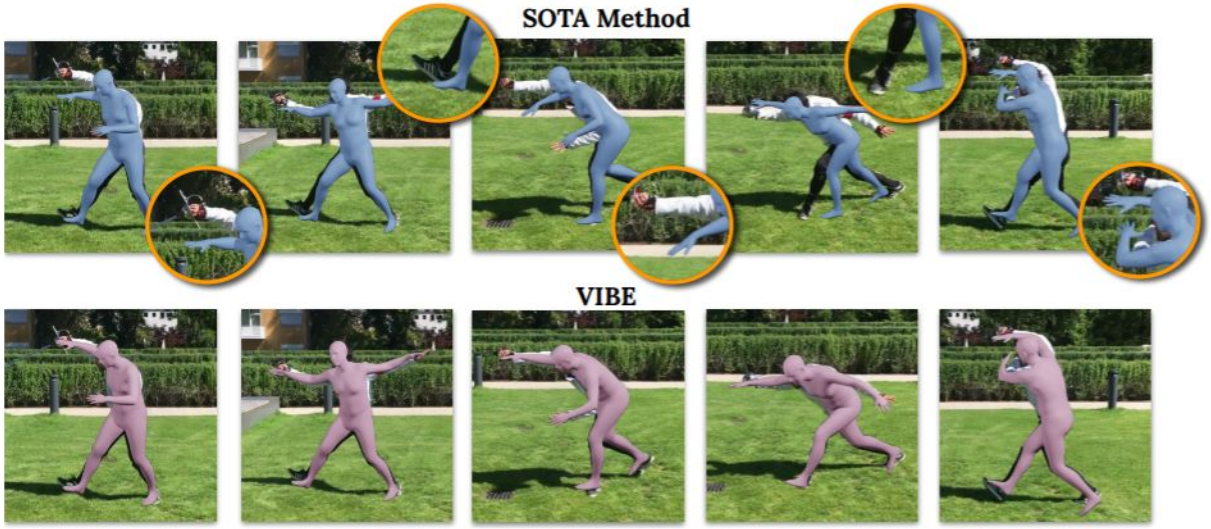


VIBE: Overall (Again)



Experiments: All Datasets

- Training
  - Human3.6M: Indoor
  - MPI-INF-3DHP: Outdoor
  - AMASS : Adversarial training
  - PennAction + PoseTrack
  - InstaVariety + Kinetics-400
- Evaluation
  - Human3.6M: Indoor
  - MPI-INF-3DHP: Outdoor
  - 3DPW: Albative study



VIBE: Quantitative Comparison

	Models	3DPW				MPI-INF-3DHP			H36M	
		PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	PCK ↑	PA-MPJPE ↓	MPJPE ↓
Frame-based	Kanazawa *et al.* [29]	76.7	130.0	-	37.4	89.8	124.2	72.9	56.8	88
	Omran *et al.* [48]	-	-	-	-	-	-	-	59.9	-
	Pavlakos *et al.* [51]	-	-	-	-	-	-	-	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.* [30]	72.6	116.5	139.3	**15.2**	-	-	-	56.9	-
	Doersch *et al.* [15]	74.7	-	-	-	-	-	-	-	-
	Sun *et al.* [58]	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

Table 1: **Benchmark of state-of-the-art models on 3DPW, MPI-INF-3DHP and H36M datasets.** Here, we compare the results of recent state-of-the-art frame-based and temporal models on 3 different datasets. VIBE(direct comp.) is our proposed model trained on video datasets similar to [30, 58]. VIBE, on the other hand, trained with extra data from 3DPW training set. VIBE outperforms all state-of-the-art models including recent well-performing SPIN [37] method on challenging 3DPW and MPI-INF-3DHP in-the-wild datasets and obtains comparable result on Human3.6M. "–" shows the results that are not available.

VIBE: Ablative Studies

	3DPW			
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓
Kanazawa *et al.* [29]	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
$\mathcal{G} + \mathcal{D}_M$ (VIBE)	**51.9**	**82.9**	**99.1**	23.4
$\mathcal{G}$ + MPoser Prior	54.1	87.0	103.9	28.2
$\mathcal{G} + \mathcal{D}_M$ + SMPLify	54.7	93.6	110.1	**27.7**

Table 2: **Ablation experiments with motion discriminator** $\mathcal{D}_M$ We experiment with several models using HMR [29] and SPIN [37] as pretrained feature extractor and add our temporal generator $\mathcal{G}$ along with $\mathcal{D}_M$. $\mathcal{D}_M$ provides consistent improvements over all baselines.

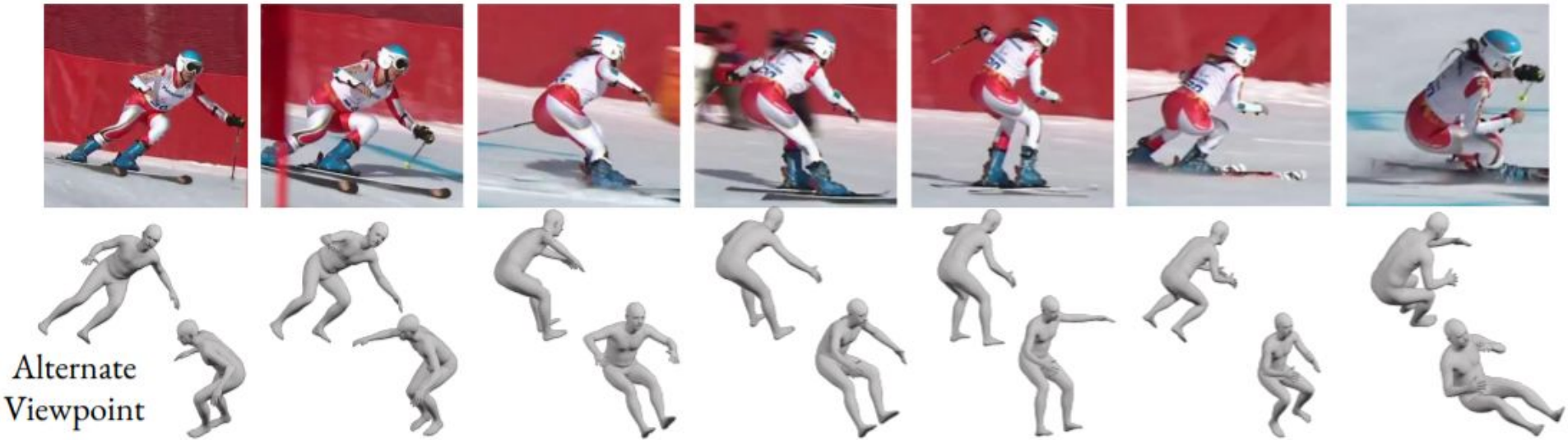
Model	PA-MPJPE ↓	MPJPE ↓
$\mathcal{D}_M$ - concat	53.7	85. 9
$\mathcal{D}_M$ - attention [2 layers,1024 nodes]	**51.9**	**82.9**
$\mathcal{D}_M$ - attention [2 layers,512 nodes]	54.2	86.6
$\mathcal{D}_M$ - attention [3 layers,1024 nodes]	52.4	82.7
$\mathcal{D}_M$ - attention [3 layers,512 nodes]	53.6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

Table 3: **Ablation experiments on self-attention** We experiment with several self-attention configurations and compare our method to static pooling approach. We report results in 3DPW dataset with different hidden sizes and numbers of layers in the MLP network that computes the attention weights.

VIBE: Qualitative Results

Qualitative comparison between VIBE (top) and Temporal-HMR (bottom).

VIBE: Qualitative Results



Conclusion

Introduced a recurrent architecture that propagates information over time;

Introduced discriminative training of motion sequences using the AMASS dataset;

Introduced self-attention in the discriminator so that it learns to focus on the important temporal structure of human motion;

Achieved SOTA result. : }