VIBE: Video Inference for Human Body Pose and Shape Estimation

Paper: https://arxiv.org/pdf/1912.05656.pdf
Dynamics - Christian Saravia

Christian Martin Saravia Hernandez

- Graduate School
  - Universidad Peruana de Ciencias Aplicadas
  - Bs. Software Engineering
- Former Job
  - Ficha Inc
    - Algorithm Development Engineer
      - Research & application of deep learning in computer vision problems
- Current Job
  - Application of machine learning / deep learning to computer vision problems
    - Object detection
    - Object classification
  - Research topics
    - Attention & Transformers

Contents

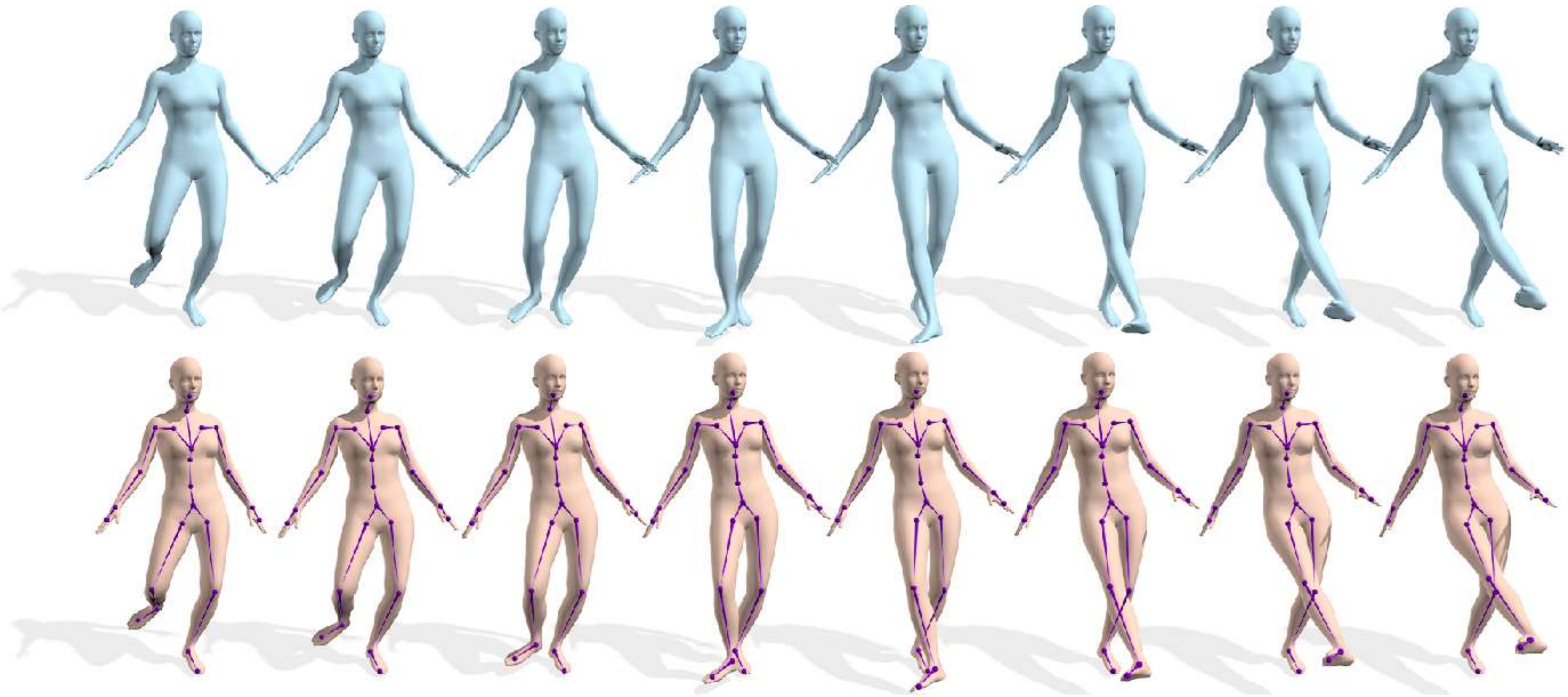
Problem

- Lack of **in-the-wild** ground-truth 3D
- Previous work combine indoor 3D datasets with videos having 2D ground-truth or pseudo ground-truth keypoint annotations
  - Indoor 3D are limited in the number of subjects, range of motion and image complexity
  - Poor amount of video labeled with ground-truth 2D pose
  - Pseudo-ground-truth 2D labels are not reliable for modeling 3D human motion

Learning 3D Human Dynamics from Video - https://arxiv.org/pdf/1812.01601.pdf

Dataset

- AMASS dataset for 3D motion capture



What is VIBE

- *"Our key novelty is an **adversarial learning framework** that leverages AMASS to **discriminate** between real human motions and those produced by our **temporal pose and shape regression networks**. We define a **novel temporal network architecture** with a **self-attention mechanism** and show that a**dversarial training**, at the sequence level, produces kinematically plausible motion sequences without in-the-wild ground-truth 3D labels."*

- **Adversarial learning framework & discriminate,** are terms used when referring to generative adversarial networks. The architecture involves the simultaneous training of two models: the generator and the discriminator. (Thanks enrico for the notes: https://www.notion.so/Generative-Adversarial-Networks-0692b1ea34e641a0ae011237345a51c4)

- ***Novel temporal network architecture.*** Since we are analyzing videos, the concept of sequence is implied. VIBE uses a gated recurrent units (GRU) to capture the sequential nature of human motion.

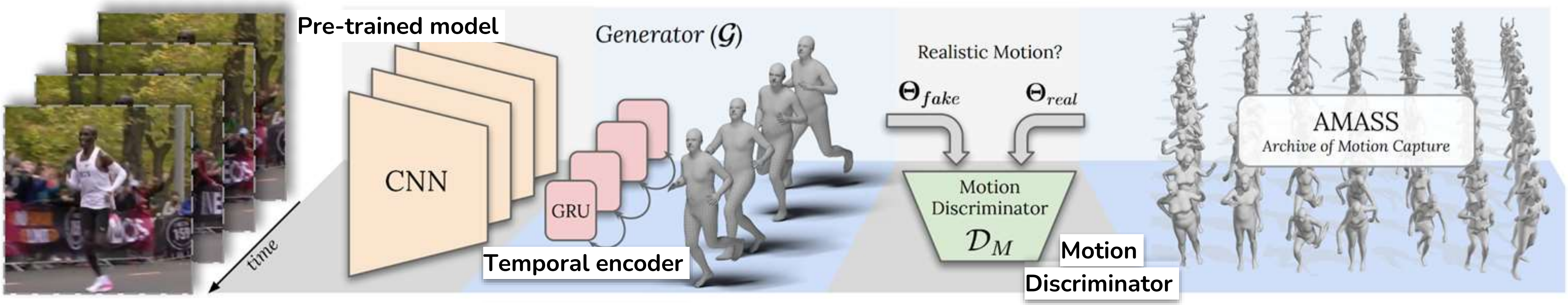
- ***Self-attention mechanism*** is used to amplify the contribution of distinctive frames.

Elements VIBE

- *“Our key novelty is an* ***adversarial learning framework*** *that leverages AMASS to* ***discriminate*** *between real human motions and those produced by our* ***temporal pose and shape regression networks****. We define a* ***novel temporal network architecture*** *with a* ***self-attention mechanism*** *and show that a****dversarial training****, at the sequence level, produces kinematically plausible motion sequences without in-the-wild ground-truth 3D labels.”*

- Architectures used:
  - Yolov3, for detecting the person box
  - Resnet50, for feature extraction
  - GRU, for sequence encoding
  - Self attention, for frame scoring
  - GAN, for adversarial training and loss

VIBE architecture



Pre-trained model

- A sequence of T frames is fed to a convolutional network, f, which functions as a feature extractor and outputs a vector $f_i \in \mathbb{R}^{2048}$ for each frame

$$f(I_1), \ldots, f(I_T)$$

Temporal encoder output

- **SMPL**

$$M(\vec{\beta}, \vec{\theta}; \Phi) \ : \ \mathbb{R}^{|\vec{\theta}| \times |\vec{\beta}|} \mapsto \mathbb{R}^{3N}$$

$\vec{\beta}$ Shape linear coefficients in a 10-dimensional space

$\vec{\theta}$ Pose vector in a 72-dimensional space

- **VIBE**

$$\hat{\Theta} = [(\hat{\theta}_1, \ldots, \hat{\theta}_T), \hat{\beta}]$$

$\hat{\beta}$ Single body shape prediction for the sequence

$\hat{\theta}_t$ Pose parameters at step t

Arithmer

Temporal encoder

- $f(I_1), \ldots, f(I_T)$ are sent to a Gated Recurrent Unit (GRU) layer that yields a latent feature vector g_i
- Then use g_i as an input to T regressors with iterative feedback.
- We use a 6D rotation representation instead of axis angles
- Loss:

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

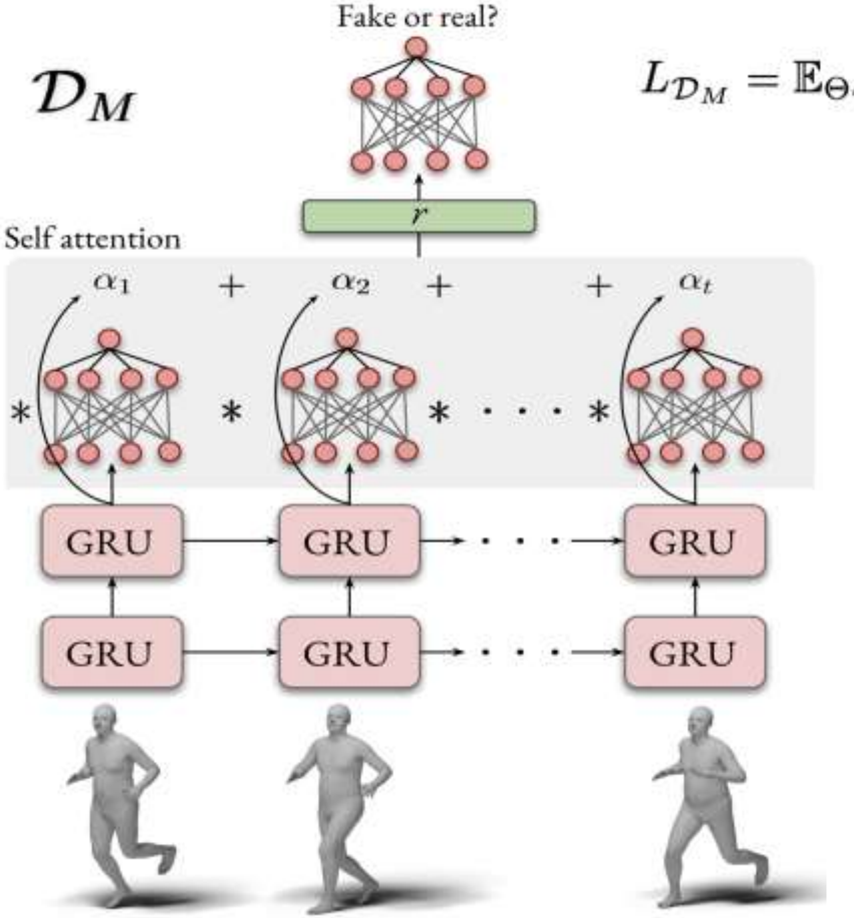
$$L_{3D} = \sum_{t=1}^{T} \|X_t - \hat{X}_t\|_2,$$

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

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

Motion Discriminator

- Enforces the generator to produce feasible real world poses that are aligned with 2D joint locations.



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

ϕ_i

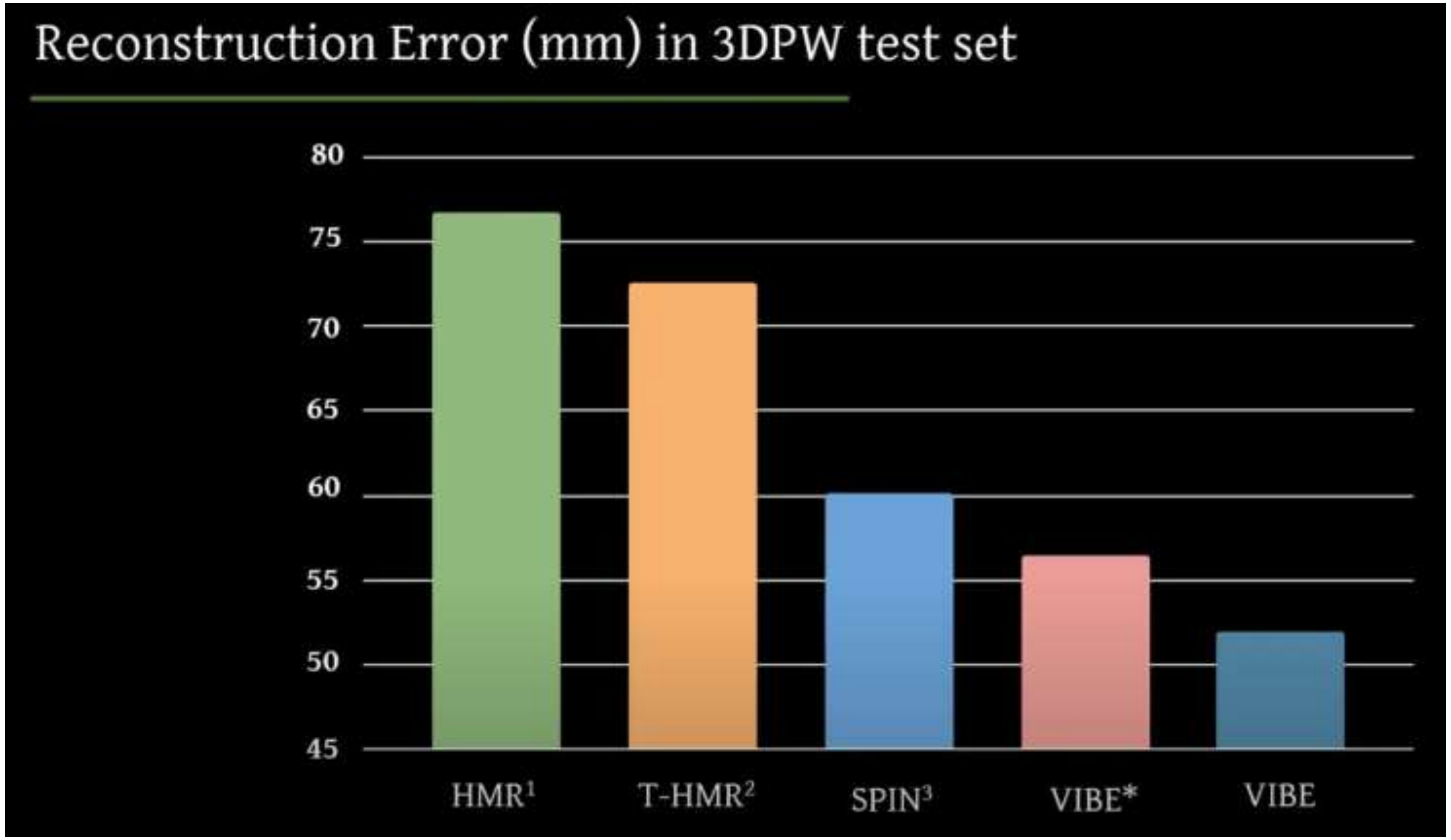
h_i

g_i

$$\phi_i = \phi(h_i), \quad a_i = \frac{e^{\phi_i}}{\sum_{t=1}^{N} e^{\phi_t}}, \quad r = \sum_{i=1}^{N} a_i h_i.$$

The weights a_i are learned by a linear MLP layer φ, and are then normalized using softmax.

Results



[1] Kanazawa et al., End-to-end Recovery of Human Shape and Pose, CVPR 2018
[2] Kanazawa et al., Learning 3D Human Dynamics from Video, CVPR 2019
[3] Kolotouros et al., Learning to Reconstruct 3D Human Pose and Shape via Modeling-fitting in the Loop, ICCV 2019

Results

Reference

- VIBE: https://arxiv.org/pdf/1912.05656.pdf
- Notes on GAN: https://www.notion.so/Generative-Adversarial-Networks-0692b1ea34e641a0ae011237345a51c4
- GAN Loss Function: https://machinelearningmastery.com/generative-adversarial-network-loss-functions/
- More of GAN: https://dl4physicalsciences.github.io/files/nips_dlps_2017_slides_louppe.pdf
- Angle to 6D Notation: https://arxiv.org/pdf/1812.07035.pdf
- Iterative Regression with 3D Feedback: https://arxiv.org/pdf/1712.06584.pdf
- Camera Weak Perspective: https://web.stanford.edu/class/cs231a/course_notes/01-camera-models.pdf