# Supplementary Material for 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
{mkocabas,nathanasiou,black}@tue.mpg.de

## 1. Implementation Details

**Pose Generator.** architecture is depicted in Figure 1. After feature extraction using ResNet50, we use 2 layer GRU network followed by a linear projection layer. The pose and shape parameters are then estimated by a SMPL parameter regressor. We employ a residual connection to assist network during training. The SMPL parameter regressor is initialized with the pretrained weights from HMR [5, 7]. We decrease the learning rate if the reconstruction does not improve for more than 5 epochs.

**Motion Discriminator.** We employ 2 GRU layers with a hidden size of 1024. For self-attention mechanism, in the case of SOTA results, we use 2 MLP layers with 1024 neurons and a dropout rate of 0.1 to estimate attention weights. For the ablation experiments we keep the same parameters changing the number of neurons and number of MLP layers only. During training, we use label smoothing for adversarial training by a random number $\in [0, 0.1]$ [10].

**Loss.** We use different weight coefficients for each term in the loss function. 2D and 3D keypoint loss coefficients are $\lambda_{2D}, \lambda_{3D} = 300$ and $\lambda_\beta = 0.06$, $\lambda_\theta = 60$. We set the motion discriminator adversarial loss term, $L_{adv}$ as $\lambda_{L_{adv}} = 2$. We use 2 GRU layers with hidden dimension size of 1024.

## 2. Datasets

Below a detailed summary of the different datasets we used for training and testing is outlined.

**MPI-INF-3DHP** [9] is a multi-view, mostly indoors dataset captured using markerless motion capture system. We use the proposed training set by authors, which consists of 8 subjects and 16 videos per subject, and we evaluate on the official test set.

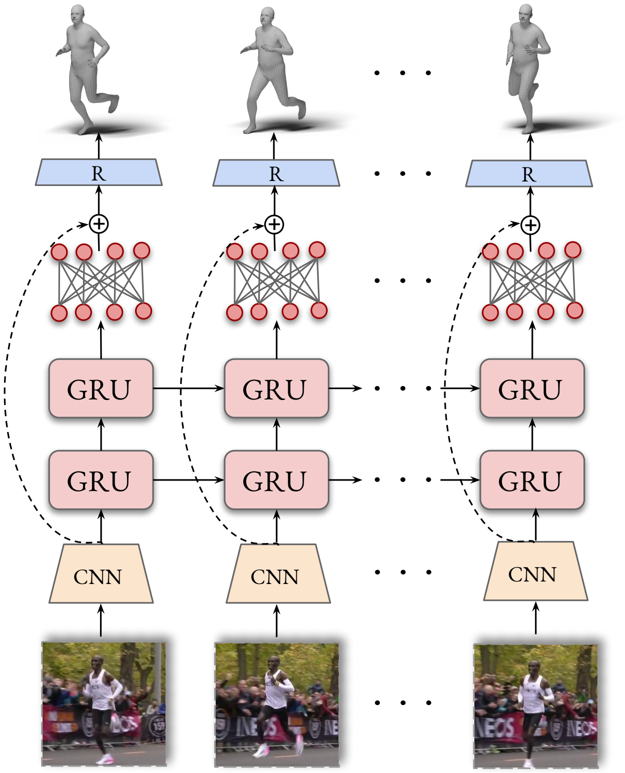

Figure 1: **Pose generator** $\mathcal{G}$ architecture used in our experiments. It takes a sequence of frames as input and output a vector $\in \mathbb{R}^{85}$

**Human3.6M** [4] Human3.6M dataset contains 15 action sequences of several individuals, captured in a controlled environment. There are 1.5 million training images with 3D annotations. We utilize SMPL parameters provided by MoSH [8] during training. Following the previous works, our model is trained on 5 subjects (S1, S5, S6, S7, S8) and tested on the other 2 subjects (S9, S11). We subsampled the dataset to 25 frames per second for training.

**3DPW** [12] a recent in-the-wild 3D dataset, captures using IMU sensors and hand-held cameras. It contains 60 videos (24 train, 24 test, 12 val) of several in-the wild and indoor activities. We use it both for evaluation and training.

**PennAction** [13] dataset contains 2326 video sequences of 15 different actions and 2D human keypoint annotations for each sequence. The sequence annotations include class label, human body joints —both 2D locations and visibility—, 2D bounding boxes and training/testing labels.

**InstaVariety** [6] is a recently curated dataset using instagram videos with particular action hashtags. It contains 2D annotations for about 24 hours of video. The 2D annotations were extracted using OpenPose [2] and Detect and Track [3] in the case of multi person scenes.

**PoseTrack** [1] PoseTrack dataset is a benchmark for multi-person pose estimation and tracking in videos. It contains 1337 videos, split into 792, 170 and 375 videos for training, validation and test set respectively. In the training split, 30 frames in the center of the video are annotated. For validation and test sets, besides the aforementioned 30 frames, every fourth frame is also annotated. The annotations include 15 body keypoints locations, a unique person id, a head and a person bounding box for each person instance in each video. We use PoseTrack during training.

## 3. Evaluation

In this section, we describe the evaluation metrics and procedures we used in our experiments. For direct comparison we used the exact same setup as in [7]. Our best results are achieved with a model that includes 3DPW training dataset in our training loop. Besides, we also get SOTA results without using it. We use Human3.6M training set when evaluating in its test set and we observe that better performance on the Human3.6M does not translate to accurate in-the-wild pose estimation.

**Metrics.** We use standard evaluation metrics for each respective dataset. First, we report the widely used MPJPE (mean per joint position error) which is calculated as the mean of the euclidean distances between the ground-truth and the predicted joint positions after aligning the pelvis. Also we use PA-MPJPE (Procrustes Aligned MPJPE) which is calculated similarly to MPJPE rigid alignment of predicted and ground-truth poses. Furthermore, we calculate Per-Vertex-Error (PVE) which is denoted by the euclidean distance between the groundtruth and predicted vertices which are the outputs of SMPL layer to demonstrate the effectiveness of VIBE. We also use the Percentage of Correct Keypoints metric (PCK) [11]. The PCK counts as correct the cases where the Euclidean distance between the actual and predicted joint positions is below a predefined threshold. Finally, we report acceleration error, that was reported in [6]. Acceleration error is the mean difference between ground-truth and predicted 3D acceleration for every joint($mm/s^2$).

## References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[3] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-Track: Efficient Pose Estimation in Videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2014. 1

[5] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[6] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[7] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, 2019. 1, 2

[8] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. In *SIGGRAPH Asia*, 2014. 1

[9] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *International Conference on 3DVision*, 2017. 1

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing*, 2016. 1

[11] Ben Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[12] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision*, 2018. 2

[13] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *International Conference on Computer Vision*, 2013. 2