Supplementary Material: Natural Human Motion Recovery by Aligning High-Order Temporal Dynamics from Monocular Videos

Dingkun Wei[1,2,*] Zehong Shen[2,*] Yan Xia[3] Georgios Pavlakos[3] Yujun Shen[2]
Xiaowei Zhou[1]
[1]Zhejiang University [2]Ant Group [3]The University of Texas at Austin

1. Overview

In this supplementary material, we provide additional implementation details of PVA-Net (Sec. 2) and extended evaluations of PVA-Net (Sec. 3), followed by a detailed description of Optimization (Sec. 4). Additionally, we demonstrate the performance improvements of HTD-Refine over the baseline on the H36M [3] dataset (Sec. 5). For a visual overview of our method and more qualitative results, please see the **supplementary video**.

2. Implementation Details of PVA-Net

Fig. 1 presents the full architecture of PVA-Net with input–output dimensions. The frame-level tokens extracted by the ViTPose encoder are fed into an 8-layer Transformer decoder that employs rotary positional embeddings (RoPE) to capture temporal dependencies across the motion sequence. We design three decoders to handle different outputs: Following the ViTPose design, our keypoint decoder employs a deconvolution layer for feature map upsampling, after which an MLP produces probability heatmaps for all 17 joints. The 2D joint locations are obtained via argmax; The velocity decoder processes sequential features through a convolutional layer followed by spatial average pooling. A Transformer module then captures temporal dynamics, after which an MLP projects the features into 3D velocity vectors for all 17 joints; The acceleration decoder adopts an identical architecture to the velocity decoder, maintaining the same processing pipeline while outputting 3D acceleration for the 17 joints.

We train PVA-Net on a mixed dataset of Human3.6M [3], BEDLAM [1], and RICH [2], normalizing targets with dataset-specific mean and standard deviation. The model is optimized with AdamW (initial learning rate 2e-4) under mixed-precision (FP16). We use a global batch size of 48 and 120-frame clips. A stepwise schedule halves the learning rate at epochs 40, 60, and 80. Training on 8×NVIDIA H20 GPUs for 100 epochs leads to convergence in approximately 48 hours. With a single A6000 GPU, PVA-Net inference operates at 30+ FPS, and requires approximately 300+ GFLOPs per frame.

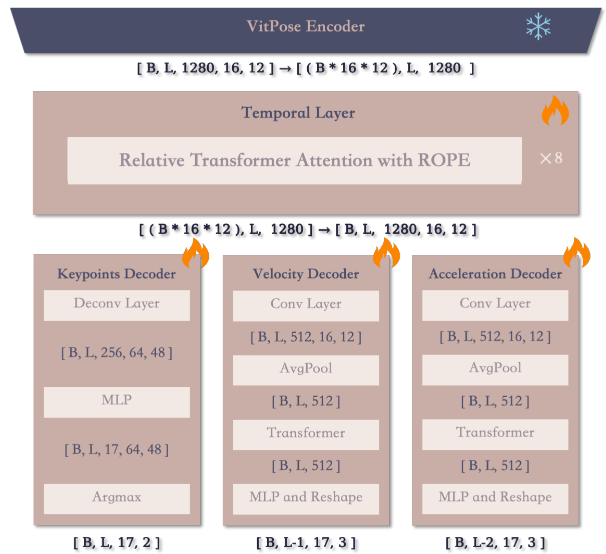

Figure 1. **Complete PVA-Net architecture with detailed layer specifications and input-output dimensions.** The system comprises: (a) ViTPose-based feature extraction, (b) 8-layer RoPE-Transformer for temporal encoding, and (c) three task-specific decoders for joint keypoints, 3D velocities, and 3D acceleration.

3. Evaluation of PVA-Net

Velocity and Acceleration. We evaluate the performance of PVA-Net on both the RICH and EMDB datasets. To comprehensively assess the accuracy of the predicted velocities and accelerations, we use the Percentage of Correct Estimates (PCE) metric under three progressively stricter thresholds.

Specifically, we define a dynamic value range $[r_{\min}, r_{\max}]$ based on the empirical 0.1 and 99.9 percentiles of the ground-truth distribution, computed from the BED-

*Equal contribution; Work done while D. Wei was an intern at Ant Group. The authors from Zhejiang University are affiliated with the State Key Lab of CAD&CG. Corresponding author: Xiaowei Zhou.

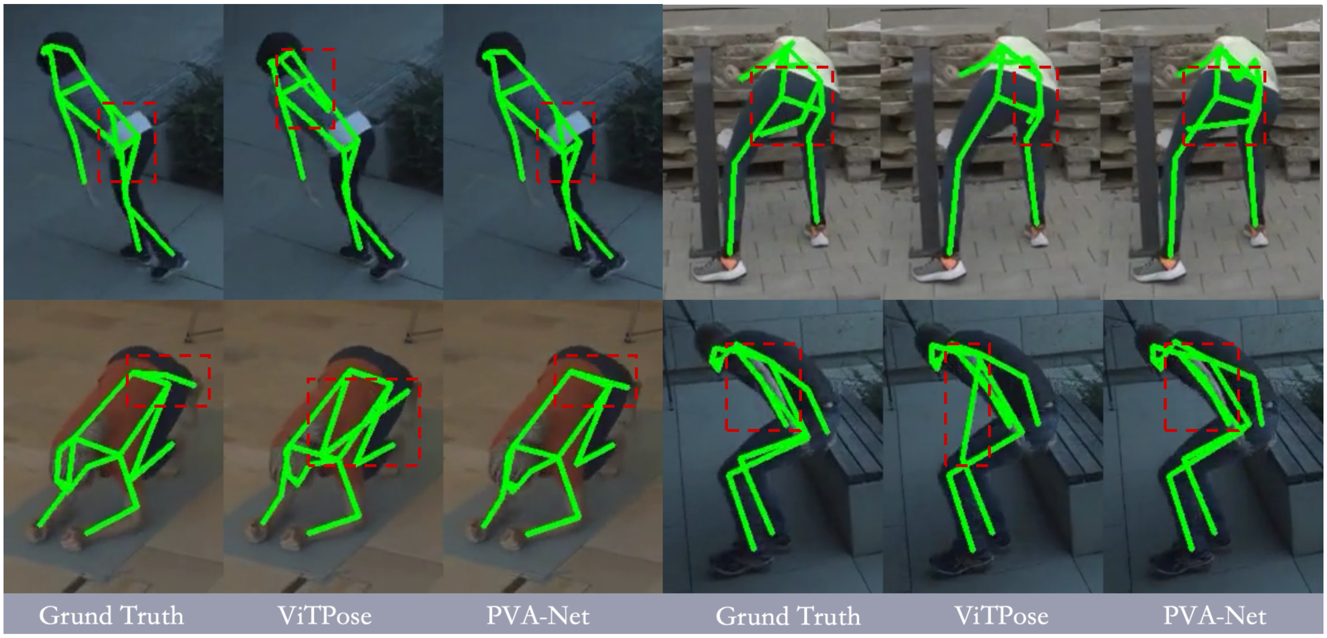

Figure 2. **Qualitative results of PVA-Net.** Our method demonstrates enhanced robustness to occlusions by effectively leveraging temporal constraints from adjacent frames to infer plausible joint positions.

Dataset	Target	PCE@0.10	PCE@0.05	PCE@0.01
EMDB	Velocity	98.2	93.0	68.2
	Acceleration	99.6	98.4	82.3
RICH	Velocity	99.9	98.7	81.9
	Acceleration	100.0	99.7	89.1

Table 1. **Velocity and Acceleration Prediction Performance.** PVA-Net achieves excellent accuracy on both EMDB and RICH datasets across PCE@0.10, PCE@0.05, and PCE@0.01 metrics, demonstrating robust performance for both velocity and acceleration estimation.

LAM [1], RICH [2] and Human3.6M [3] datasets:

$$r_{\min} = \text{Percentile}_{0.1}(\mathcal{D}), \qquad r_{\max} = \text{Percentile}_{99.9}(\mathcal{D}), \tag{1}$$

where $\mathcal{D}$ denotes the ground-truth velocity/acceleration statistics. A prediction is considered correct if its absolute error is below 10%, 5%, or 1% of this dynamic range, corresponding to PCE@0.10, PCE@0.05, and PCE@0.01, respectively. For completeness, the PCE metric at threshold τ is defined as:

$$\text{PCE@}\tau = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(|\hat{\mathcal{D}}_i - \mathcal{D}_i| < \tau\left(r_{\max} - r_{\min}\right)\right), \tag{2}$$

where $\mathcal{D}_i$ and $\hat{\mathcal{D}}_i$ denote the ground-truth and predicted values, and N is the number of evaluations.

As shown in Tab. 1, PVA-Net demonstrates strong performance across both datasets and evaluation metrics. Several key observations emerge: First, acceleration predictions consistently outperform velocity predictions, achieving higher accuracy scores, which aligns with our emphasis on acceleration modeling in the main paper. Secondly, we observe a performance gap between RICH and EMDB datasets (better on RICH). This discrepancy primarily stems from the inherent coupling of camera and human motion in EMDB, which presents additional challenges for estimation. These results validate that PVA-Net effectively learns discriminative velocity and acceleration representations for 3D human pose dynamics.

Keypoint. In our keypoint detection comparison, to ensure a fair evaluation, we fine-tune ViTPose-L [6] on BEDLAM, RICH, and Human3.6M datasets. We employ two primary metrics for performance assessment: PCK@10 and PCK@05, which measure the percentage of correct keypoints within 10 pixels and 5 pixels of ground truth positions, respectively. Additionally, to demonstrate the temporal stability of our method, we evaluate the acceleration error (ACCEL) between predicted 2D keypoints and ground truth annotations.

The experimental results presented in Tab. 2 demonstrate the superior performance of our method across multiple datasets. Our approach achieves higher PCK accuracy while reducing acceleration errors, indicating improvements in both spatial precision and temporal consistency. This perfor-

Dataset	Method	PCK@10	PCK@05	ACCEL
EMDB	ViTPose-L (w/o refine)	95.5	78.8	4.9
	ViTPose-L (w/ refine)	97.0	85.0	4.5
	PVA-Net	98.3	88.8	3.7
RICH	ViTPose-L (w/o refine)	91.1	71.5	2.0
	ViTPose-L (w/ refine)	94.5	83.1	2.0
	PVA-Net	97.0	87.7	0.9

Table 2. **2D Keypoint Detection Performance on EMDB and RICH.** Evaluation on EMDB and RICH datasets shows that our PVA-Net achieves the best performance across all metrics (PCK@10, PCK@05, ACCEL).

mance gain is primarily due to the temporal layer architecture that incorporates temporal constraints, resulting in more stable predictions.

Additionally, Fig. 2 further demonstrates that our method effectively resolves occlusion-induced ambiguities through inter-frame temporal reasoning. In the **supplementary video**, we provide additional qualitative results showing that PVA-Net achieves substantially improved stability and accuracy compared to ViTPose-L across diverse motion scenarios.

4. Implementation Details of Optimization

We optimize the global SMPL parameters end-to-end using the Adam optimizer with an initial learning rate of 1e-3 for 1500 epochs. The learning rate is linearly warmed up during the first 10 epochs to stabilize training, and reduced by a factor of 10 at epoch 1000. With our fully parallelized implementation, the optimization for a single video completes in approximately 2 minutes on an NVIDIA A6000 GPU, and the runtime remains nearly constant regardless of video length, since all frames are optimized jointly in a batched formulation.

5. Evaluation of HTD-Refine on H36M

We benchmark HTD-Refine using representative HMR methods, TRAM [5] and GVHMR [4], on the H36M dataset [3]. As shown, HTD-Refine consistently improves performance

H36M Results	Jitter	FS	MPJVE	MPJAE	WA-MPJPE	W-MPJPE	RTE
TRAM (w/ traj filter)	10.4	8.4	0.3	4.9	77.7	141.7	1.5
TRAM+HTD-Refine	**4.0**	**2.9**	**0.2**	**3.1**	**69.8**	**121.4**	**1.4**
GVHMR	10.9	2.1	0.2	4.7	51.1	84.3	1.4
GVHMR+HTD-Refine	**3.6**	**2.0**	0.2	**2.7**	**44.4**	**68.6**	**0.8**

Table 3. **Quantitative comparison on the H36M test set.**

for both TRAM and GVHMR across all metrics. Specifically, it significantly reduces temporal jitter and global positioning errors, while also suppressing artifacts like foot sliding. These results are consistent with our experiments on EMDB and RICH, further validating the effectiveness and plug-and-play capability of our module.

6. Limitations and Future Work

While HTD-Refine demonstrates effective enhancement over existing methods, we acknowledge several challenges for future work. First, our approach relies on the outputs of camera and human pose estimation, which can introduce errors into the optimization process. Second, the generalization ability of PVA-Net is limited in extreme scenarios underrepresented in the training data, such as skateboarding. In addition, under severe occlusion or motion blur, decreased confidence of PVA-Net causes the optimization to rely on regularization priors, tending to retain the initialization. Finally, although our method jointly optimizes full video sequences, the computational overhead is still significant enough to limit real-time use. Future work may explore more efficient or streaming-based solutions.

References

[1] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023-06. 1, 2

[2] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022-06. 1, 2

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1, 2, 3

[4] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024. 3

[5] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *Eur. Conf. Comput. Vis.*, 2024. 3

[6] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 2