—Supplementary Materials— Kinematic-aware Hierarchical Attention Network for Human Pose Estimation in Videos

Kyung-Min Jin[1], Byoung-Sung Lim[1], Gun-Hee Lee[2], Tae-Kyung Kang[1], and Seong-Whan Lee[1]

[1]Department of Artificial Intelligence, Korea University
[2]Department of Computer Science and Engineering, Korea University
{km_jin, bs_lim, gunhlee, tk_kang, sw.lee}@korea.ac.kr

We further elaborate descriptions of datasets [4, 3, 1, 2, 12, 7] and off-the-shelf estimators [5, 10, 6, 13, 8] used in our training. In Sec. 2, we show more ablation studies on 3DPW [12], highlighting the effect of the proposed keypoint kinematic features, which significantly impact learning temporal relationships between consecutive frames' poses, as the main paper (ablation studies on Sub-JHMDB [4]). Finally, in Sec. 3, we present the additional qualitative results and comparison on 2D/3D pose estimation and verify the effectiveness of HANet. We strongly recommend watching our supplementary video at `https://youtu.be/pe3okGjj7mo` that describe existing pose estimators, which often suffer severe jitter in highly-occluded scenes for both 2D/3D pose estimation contrary to HANet. Code is available at `https://github.com/KyungMinJin/HANet`.

1. Datasets

Sub-JHMDB. Sub-JHMDB [4] is a video 2D human pose dataset, which is a subset of JHMDB. It contains 316 videos, and the average duration is 35 frames. For each frame, it provides 15 annotated body keypoints. We use the bounding box calculated from [9] and mix three original splitting schemes for training and testing in 2D pose estimation experiments, following [11, 15, 14]. We use SimpleBaseline [13] as an input pose estimator that provides a simple and effective baseline method.

PoseTrack[1] PoseTrack2017 [3] and PoseTrack2018 [1] are sparsely annotated large-scale benchmark datasets for multi-person pose estimation and tracking in videos. PoseTrack2017 contains 514 videos with 66,374 frames. PoseTrack 2018 increased the number of videos to 1,138 with 153,615 pose annotations. Thirty frames from the center are annotated for training videos. Conversely, every fourth frame is annotated for validation and test videos. The annotations include 17 body keypoints locations with visibility, a unique person id, and a head bounding box for each person instance. Our model uses DCPose [8], which is state-of-the-art for PoseTrack, by refining the current frame's pose in the video using previous and next frames' poses.

Component				*MPJPE*	*Accel*
Flow	Vel.	Accel.	WB		
				77.2	8.6
✓				76.9	8.5
✓	✓			76.4	8.3
✓	✓	✓		75.7	8.1
✓	✓	✓	✓	**74.6**	**8.0**

Table 1. Ablation study on keypoint kinematic features. We report *MPJPE* and *Accel* errors on 3DPW [12] dataset. WB stands for weight and bias of velocity and acceleration.

Human3.6M. Human3.6M [2] is a large-scale indoor video dataset with 15 actions from 4 camera viewpoints. It has 3.6 million frames at 50 frames-per-second (fps). 3D human joint positions are captured accurately from a high-speed motion capture system. Following previous works [14, 10], we use the standard protocol with 5 actors (S1, S5, S6, S7, S8) as the training set and another 2 actors (S9, S11) as the testing set. We leverage FCN [10] that uses multiple fully connected layers along the spatial dimension.

3DPW. 3DPW [12] is a challenging in-the-wild dataset consisting of more than 51k frames at 30 fps with accurate 3D poses and shapes annotation. This dataset is used to validate the performance of body mesh recovery methods. We use PARE [5] for pose estimator which handles partial occlusion scenes with a part-guided attention mechanism.

[1] We only report the evaluation results on validation sets in the main paper because we could not evaluate our HANet on test sets due to the service outage with posetrack homepage `https://posetrack.net`.

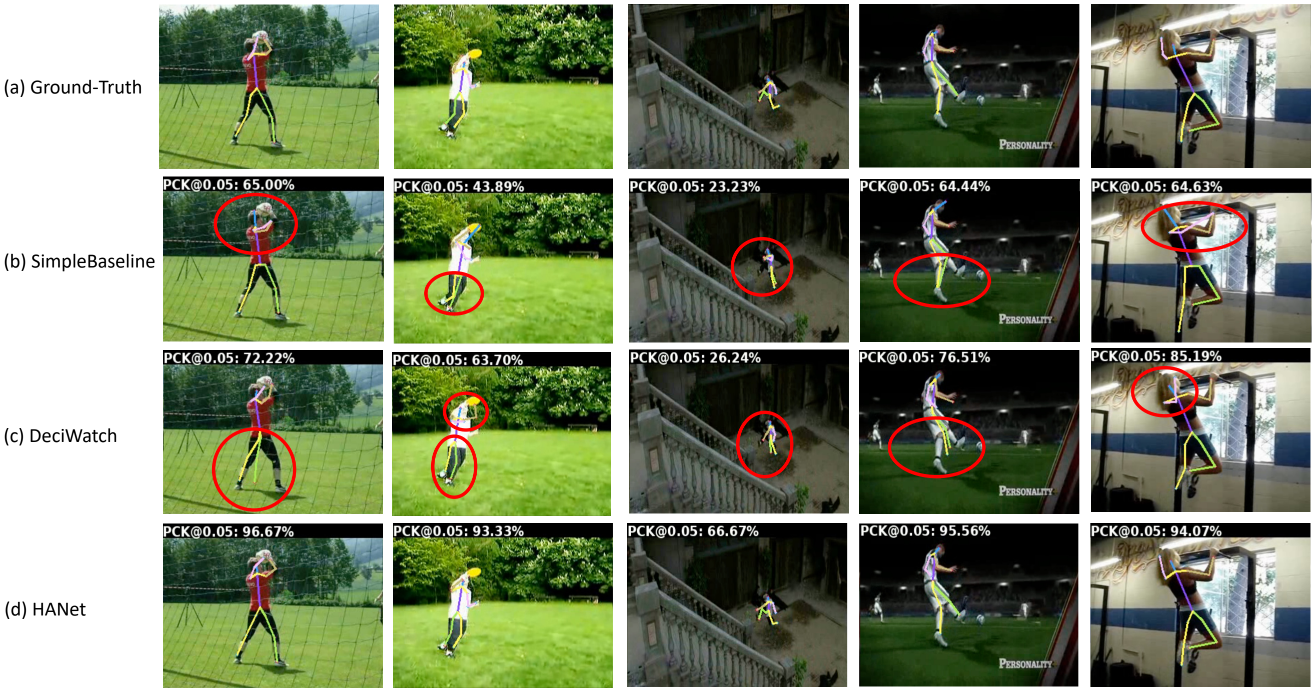

Figure 1. Visualization of 2D pose estimation with state-of-the-art method [14] and input pose estimator [13] on videos from Sub-JHMDB [4] dataset.

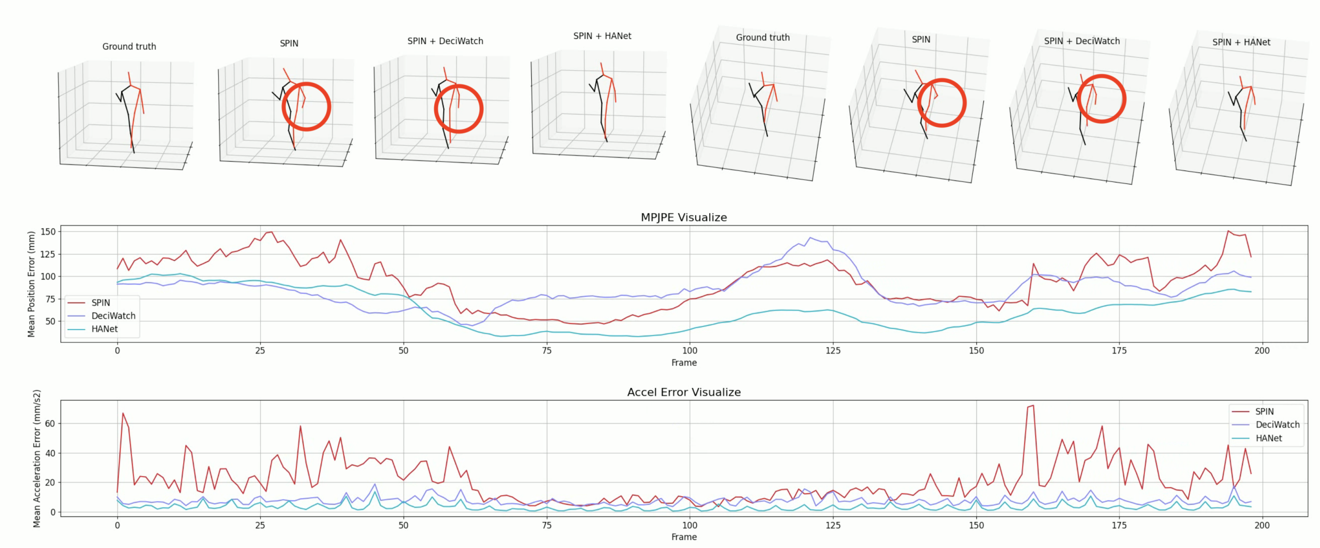

Figure 2. *MPJPE* and *Accel* error comparison with prior works [10, 14] on a video from AIST++ [7] dataset.

AIST++. AIST [7] is a challenging dataset with diverse and fast-moving dances that comes from the AIST Dance Video DB. It contains 3D human motion annotations of 1, 408 video sequences at 60 fps, which is 10.1M frames in total. With 3D human keypoint annotations and SMPL parameters, it covers 30 different actors in 9 views. We use SPIN [6] that optimize SMPL to regress body shape and pose parameters, as our off-the-shelf methods to refine 3D poses or body meshes.

2. Additional Ablation Study

To verify the effectiveness of our keypoint kinematic features on 3D pose estimation and body mesh recovery, we conduct additional ablation studies on 3DPW [12] dataset. As shown in Table 1, we demonstrate that keypoint kinematic features impact on performances, significantly drop *MPJPE* and *Accel* errors both, which means we can accurately produce smooth 3D poses with these features.

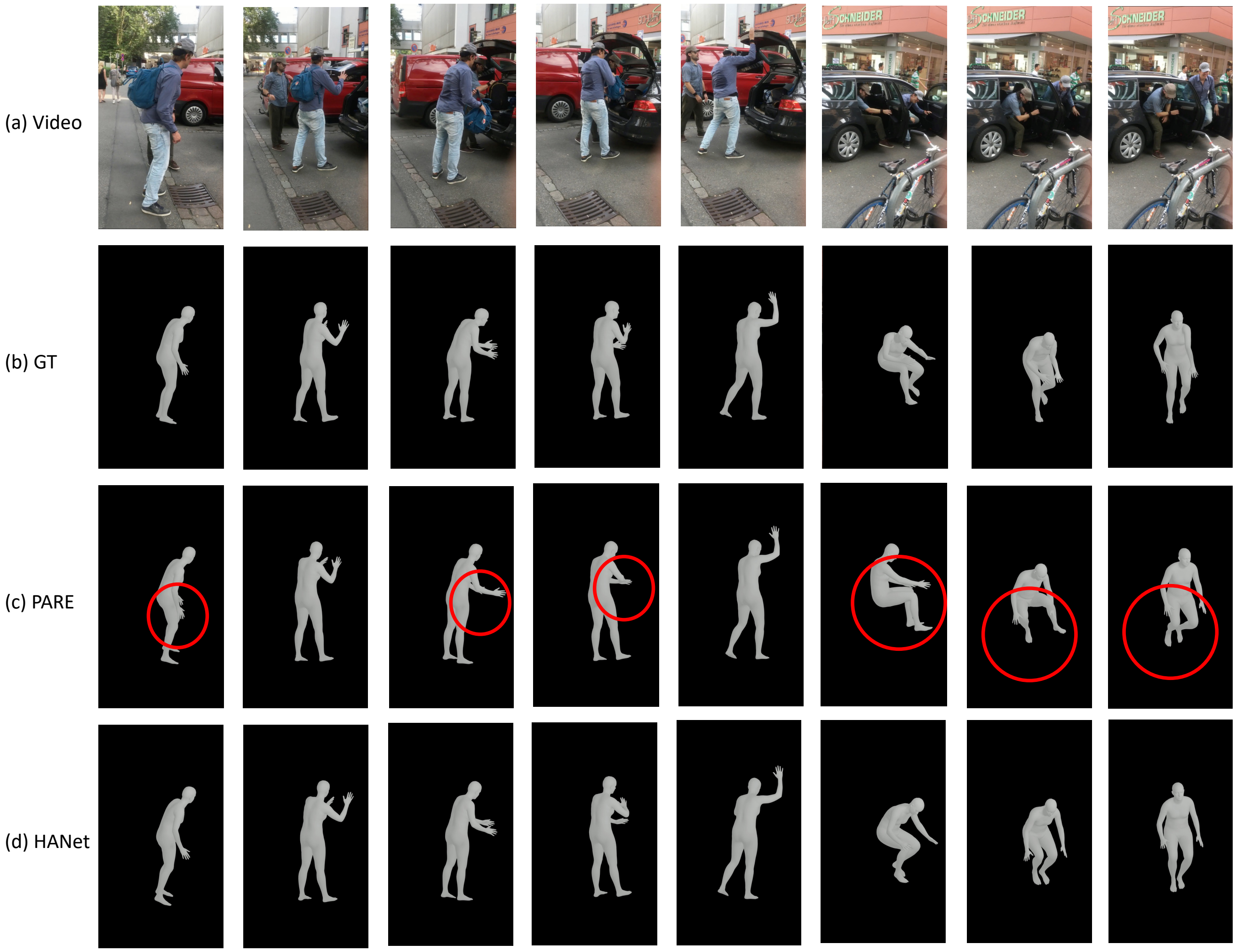

Figure 3. Visualization of 3D body mesh recovery with state-of-the-art input pose estimator [5] on a video from 3DPW [12] dataset.

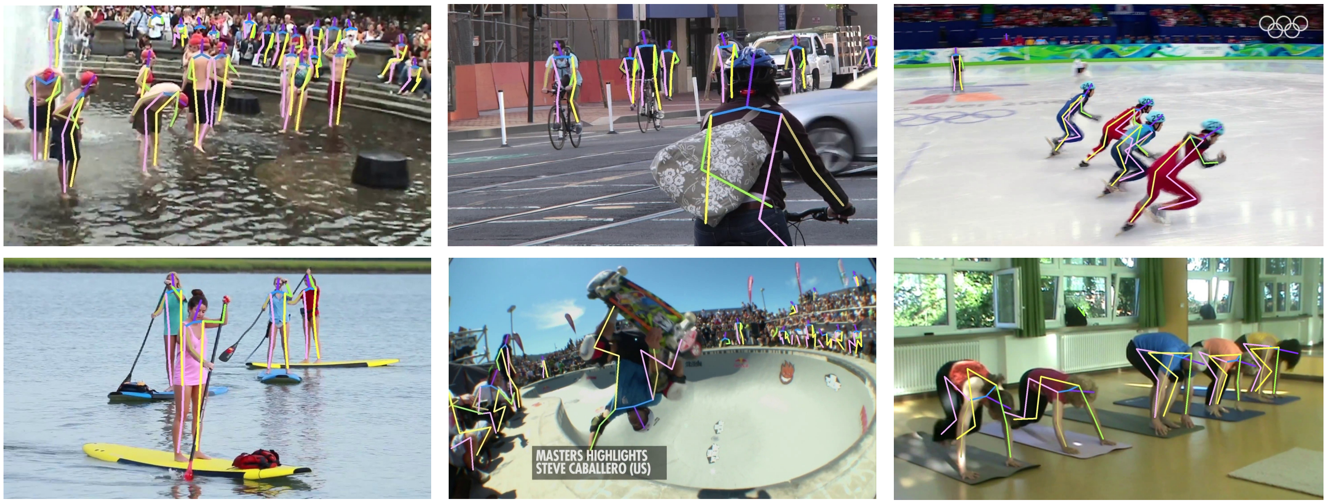

Figure 4. Visualization of multi-human pose estimation on videos from sparsely-annotated PoseTrack [3] dataset.

3. Additional Experiments

We conduct additional experiments and visualize qualtative results and comparisons in Fig. 1 - Fig. 4. As illustrated in Fig. 1, SimpleBaseline and existing state-of-the-art DeciWatch on Sub-JHMDB shows inaccurate results in such cases: flipped left and right symmetry, occlusion, and motion blur. The red circles show the wrong estimated parts. In Fig. 2, we show estimated 3D poses from input pose estimator [6], ours, and DeciWatch [14] in two camera-views. Contrary to our model, [6, 14] do not accurately estimate the position of fast-moving body parts, such as right arm. We demonstrate lower *MPJPE* and *Accel* errors shown as graphs in Fig. 2.

Moreover, we visualize 3D body mesh on a video from 3DPW [12] datasets. Compared to the state-of-the-art body mesh recovery method [5], HANet accurately recovers 3D body mesh and mitigate jitter by learning the temporal relationship of consecutive poses through the kinematic characteristics of keypoints.

Lastly, we show the muti-human 2D pose estimation results on PoseTrack2017 [3] and PoseTrack2018 [1]. In Fig. 4, we visualize the accurate results on crowded, highly-occluded, motion blur, and challenging poses, and verify that our HANet can be effectively applied to multi-human pose estmation and learn temporal relationships between poses without full-supervision.

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 *CVPR*, 2018.

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2013.

[3] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *CVPR*, 2017.

[4] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, 2013.

[5] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021.

[6] 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 *ICCV*, 2019.

[7] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021.

[8] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *CVPR*, 2021.

[9] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. In *CVPR*, 2018.

[10] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[11] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *ICCV*, 2019.

[12] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018.

[13] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.

[14] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for 10x efficient 2d and 3d pose estimation. In *ECCV*, 2022.

[15] Yuexi Zhang, Yin Wang, Octavia Camps, and Mario Sznaier. Key frame proposal network for efficient pose estimation in videos. In *ECCV*, 2020.