Supplementary Material: Just Add π! Pose Induced Video Transformers for Understanding Activities of Daily Living

Dominick Reilly
UNC Charlotte
dreilly1@charlotte.edu

Srijan Das
UNC Charlotte
sdas24@charlotte.edu

In this supplementary material, we provide the following additional details:

- Motivation for the use of Pose over Flow.
- Datasets and Evaluation Protocols
- Additional Implementation Details
- The Effects of Noisy Poses
- Improvement Cases
- Comparison with baseline and 3D skeleton model
- Feature Analysis of 2D-SIM

A. Motivation for the use of Pose over Flow.

While optical flow is useful for web video datasets where actions are defined by prominent motion, such as in Kinetics [8] and AVA [7], it is not effective on ADL datasets where the motion cues are more subtle [3, 4]. A majority of approaches that distil optical flow into the RGB stream are only evaluated web video datasets. In Table 1, we present results on Smarthome using optical flow. To demonstrate pose's superiority over optical flow for ADL, we distill optical flow into an RGB model following MARS [2]. For a fair comparison with our methods, we also adapt the CNN backbone in MARS with TimeSformer, denoted as MARS*.

Table 1. Justification for use of Pose over Optical Flow.

Smarthome	RGB only	Pose Only	Flow Only	MARS	MARS*	π-ViT
CS	68.4	57.5	51.8	58.1	60.7	**72.9**
CV2	60.6	35.2	34.1	45.7	56.5	**64.8**

B. Datasets and Evaluation Protocols

We evaluate our methods on three popular Activities of Daily Living (ADL) datasets.

Toyota-Smarthome [3] (Smarthome, SH) provides 16.1K video clips of elderly individuals performing actions in a real-world smarthome setting. The dataset contains 18 subjects, 7 camera views, and 31 action classes. For evaluation, we follow the cross-subject (CS) and cross-view (CV$_1$, CV$_2$) protocols. Due to the unbalanced nature of the dataset, we use the mean class-accuracy (mCA) performance metric. The dataset provides 2D and 3D skeletons containing 13 keypoints that were extracted using LCRNet [12], which are used to generate the inputs to our 2D-SIM and 3D-SIM approaches.

NTU120 [9] provides 114K video clips of subjects performing actions in a controlled laboratory setting. The dataset consists of 106 subjects, 3 camera views, and 120 action classes. We follow the cross-subject (CS) and cross-setup (CSet) protocols for evaluation, and report the top-1 classification accuracy. The dataset provides 2D and 3D skeletons containing 25 keypoints extracted using Microsoft Kinect v2 sensors, which we use to generate the inputs to our 2D-SIM and 3D-SIM approaches.

NTU60 [13] is a subset of NTU120 that provides 56.8K video clips of subjects performing actions in a controlled laboratory setting. The dataset consists of 40 subjects, 3 camera views, and 60 action classes. For evaluation, we follow the cross-subject (CS) and cross-view (CV) protocols, and report the top-1 accuracy. For our ablations, we follow the cross-view-subject (CVS) protocol, CVS1, as proposed in [14]. In the CVS protocols, the subjects and viewpoints in the training set are distinct from the subjects and viewpoints in the testing set. Specifically, only the 0° viewpoint from the NTU60 CS training protocol is used for training, while testing is carried out on the 0°, 45°, or 90° viewpoints from the NTU60 CS test split, which are referred to as CVS1, CVS2, and CVS3, respectively. We use the CVS protocols because they provide a better represent the cross-view challenge. The dataset provides 2D and 3D skeletons containing 25 keypoints extracted using Microsoft Kinect v2 sensors, which we use to generate the inputs to our 2D-SIM and 3D-SIM approaches.

C. Implementation Details (Additional)

We train all our models on 8 RTX A5000 or A6000 GPUs.

Our models. In all experiments, we use a 12 layer TimeSformer [1] video transformer backbone and follow a training pipeline similar to [1]. We use Kinetics400 pretraining for Smarthome and SSv2 pretraining for NTU60 and NTU120. For fine-tuning, we train our models for 15 epochs. The

RGB inputs to our models are video frames of size 8 × 224 × 224 for Smarthome and a size of 16 × 224 × 224 for NTU60 and NTU120. Frames are sampled at a rate of $\frac{1}{32}$ for Smarthome and uniform sampling is used for NTU60 and NTU120. As done in [4, 5], we extract 224 × 224 human crops from the video before feeding them to our models. This ensures that the video frames input to our model will contain human skeleton joints.

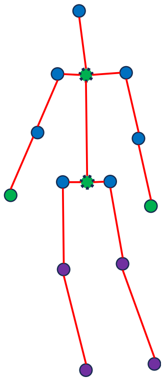


Figure 1. Illustration of the human joint partitions used when training Hyperformer on Smarthome. Color denotes partition, dashed outline indicates interpolated human joint.

Other video transformers. For results we generate ourselves using other video transformer methods [10, 11] (indicated by † in SoTA tables), we follow the default configurations suggested by each method. For a fair comparison with our models, we also utilize Kinetics400 pretraining for Smarthome and SSv2 pretraining for NTU60 and NTU120.

3D skeleton model. As mentioned in the main paper, we use Hyperformer [15] as the pretrained 3D skeleton model in 3D-SIM. For Smarthome, we train Hyperformer using the human joint partition shown in Figure 1, otherwise we follow the same training configuration proposed in [15]. Note that we interpolate the base-of-spine and top-of-spine keypoints. This allows the origin of the joints to be centered at the spine, a required pre-processing step in Hyperformer.

D. The Effects of Noisy Poses

Figure 2 highlights the challenges encountered in real-world human pose estimation, particularly evident in the Smarthome dataset. Issues such as occlusions and unusual camera angles frequently degrade the accuracy of pose estimations in such settings. It is worthwhile to mention that datasets like NTU tend to exhibit fewer of these complications due to their controlled collection environments and use of specialized sensors for collecting poses, both of which are impractical in the real world.

Table 2. Top-5 classes improved by 2D-SIM and 3D-SIM on Toyota-Smarthome CS and NTU120 CS.

(a) 2D-SIM

Action name	**Improvement over baseline**
Toyota-Smarthome	
Drink.FromGlass	+33.3%
Use Tablet	+13.3%
Drink.Frombottle	+11.4%
WatchTV	+10.0%
MakeCoffee.PourGrains	+9.5%
NTU120	
Cut Paper w/ Scissors	+4.0%
Reading	+2.9%
Thumb down	+2.4%
Shoot at basket	+2.3%
Clapping	+2.2%

(b) 3D-SIM

Action name	**Improvement over baseline**
Toyota-Smarthome	
Eat Snack	+13.7%
Maketea.Boilwater	+12.5%
Cook.Usestove	+11.1%
Pour.Frombottle	+10.6%
WatchTv	+10.0%
NTU120	
Rub hands together	+6.9%
Make victory sign	+5.0%
Wield knife at person	+4.5%
Yawn	+4.4%
Play magic cube	+3.7%

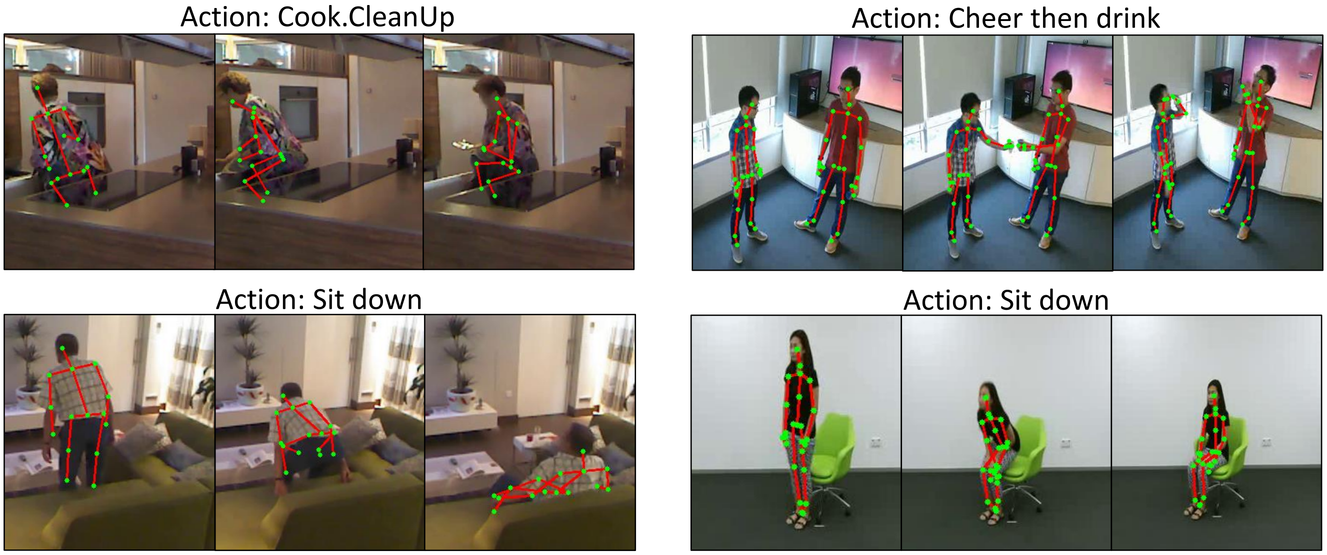
These observations exemplify the need to design algorithms that are robust to noisy poses. In Figure 3, we introduce varying levels of noise into 2D-SIM and 3D-SIM to evaluate their effectiveness in the presence of noisy poses. For 2D-SIM, we directly introduce pixel-level noise into the human skeleton joints. For each joint coordinate, we randomly sample two values between 0 and the designated noise level and add it to the joints x and y coordinates. For 3D-SIM, we add noise to the 3D skeleton features used as input to the model. The levels of noise chosen are based on the standard deviations of the features, and then, similarly to 2D-SIM, we randomly sample values between 0 and the designated noise level and add it to the feature vector to generate noisy 3D skeleton features.

In Figure 3a, we observe that 2D-SIM is sensitive to very noisy poses, causing the performance to match the baseline. This makes sense as with wildly inaccurate poses, the extra supervision provided by 2D-SIM will be wasted on non-salient RGB regions. At low to medium levels of noise (20, 40), which are more likely to apply in real-world pose estimation, 2D-SIM still provides improvements over the baseline. Figure 3b shows the effect of noisy 3D skeleton features on 3D-SIM. We see that 3D-SIM is more robust to high levels of noise, consistently outperforming the baseline across all noise levels. This can be attributed to the inherent robustness of 3D skeleton models to noisy poses, as shown in previous research [6].

E. Improvement Cases

In Table 2, the top-5 action classes demonstrating significant performance enhancements via 2D-SIM and 3D-SIM over the baseline in the Toyota-Smarthome CS and NTU120 CS protocols are presented. Notably, the largest improvements of 2D-SIM are observed in actions with fine-grained appearance details, such as *Drink from glass* and *Use Tablet*. This indicates the effectiveness of 2D-SIM on actions where modeling fine-grained appearance is necessary. For 3D-SIM, the largest improvements come from actions with fine-grained motion, e.g., *rub hands together*.

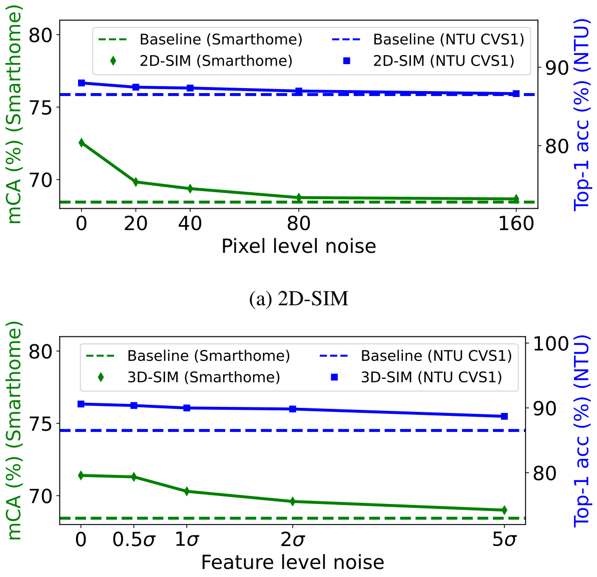
In Table 3 and Table 4, we present the the top-5 class-



(a) Sample videos and poses from Toyota-Smarthome.

(b) Sample videos and poses from NTU.

Figure 2. Visualizations of poses from Smarthome (a) and NTU (b).



(a) 2D-SIM

(b) 3D-SIM

Figure 3. Effects of noisy poses on 2D-SIM and 3D-SIM on Smarthome CS and NTU CVS1 protocols.

pairs that are improved by 2D-SIM and 3D-SIM on the Toyota-Smarthome CS and NTU120 CS protocols. The metric displayed is the raw number of predictions, i.e., the number Action 1 samples that were misclassified as Action 2. We observe that the baseline often confuses actions with similar appearance, such as confusing *Takepills* with *UseTelephone* or *Drink.Frombottle* with *Drink.Fromglass*.

Table 3. Top-5 class-pairs improved by 2D-SIM over the Baseline (TimeSformer) on Toyota-Smarthome CS and NTU120 CS.

Action 1	**Action 2**	**2D-SIM Improvement over baseline**
	Toyota-Smarthome	
Takepills	UseTelephone	+61.5%
Pour.Frombottle	Pour.Fromcan	+60.0%
Drink.Frombottle	Drink.Fromglass	+57.1%
WatchTV	ReadBook	+40.6%
Drink.FromCan	WatchTV	+29.4%
	NTU120	
Toss coin	Make ok sign	+38.9%
Reading	Writing	+23.5%
Make victory sign	Make ok sign	+14.3%
Yawn	Blow nose	+13.6%
Cut Paper w/ Scissors	Staple book	+12.7%

Table 4. Top-5 class-pairs improved by 3D-SIM over the Baseline (TimeSformer) on Toyota-Smarthome CS and NTU120 CS.

Action 1	**Action 2**	**3D-SIM Improvement over baseline**
	Toyota-Smarthome	
WatchTV	UseTelephone	+42.85%
WatchTV	ReadBook	+33.33%
Cook.Cleanup	Walk	+29.41%
Cook.Cleandishes	Cook.Cleanup	+15.15%
Enter	Leave	+7.10%
	NTU120	
Yawn	Flick hair	+54.55%
Rub hands together	Clapping	+32.43%
Yawn	Blow nose	+25.76%
Make victory sign	Make okay sign	+25.51%
Cut paper	Staple book	+13.64%

We also observe that 2D-SIM can improve performance in such cases, owing to the additional supervision applied to the salient RGB regions. We also observe that the baseline confuses actions with similar motion, such as *Yawn* vs *Blow nose*, and show that 3D-SIM improves the performance in these cases.

Table 5. Comparison of our methods to the 3D skeleton model used in 3D-SIM (Hyperformer) and the baseline TimeSformer.

Method	Toyota-Smarthome			NTU60		NTU120	
	CS	**CV$_1$**	**CV$_2$**	**CS**	**CV**	**CS**	**CSet**
Hyperformer [15]	57.5	31.6	35.2	90.7	95.1	86.6	88.0
π-ViT + 3D Poses	73.1	55.6	65.0	96.3	99.0	95.1	96.1
TimeSformer [1]	68.4	50.0	60.6	93.0	97.2	90.6	91.6
+ 2D-SIM	72.5	54.8	62.9	93.0	97.0	90.5	91.6
+ 3D-SIM	71.4	51.2	62.3	94.0	97.8	91.8	92.7
π-ViT	72.9	55.2	64.8	94.0	97.9	91.9	92.9

F. Comparison with baseline and 3D skeleton model

In Table 5, we compare our methods with the baseline TimeSformer [1], our video transformer backbone, and with Hyperformer [15], the 3D skeleton model used in 3D-SIM. We first observe the disparity in performance on Smarthome between Hyperformer and TimeSformer, owing this to the noisy poses in Smarthome and the importance of appearance in distinguishing the actions. This is further evident from the relatively small improvement seen on Smarthome when adding 3D poses to π-ViT compared to NTU60 and NTU120. On the NTU datasets, we observe that the TimeSformer outperforms Hyperformer, but that both modalities are quite complementary (as evidenced by π-ViT + 3D Poses).

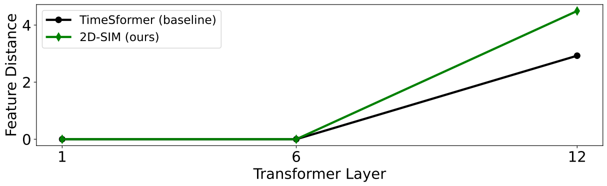


Figure 4. Average feature distance between pose tokens.

G. Feature Analysis of 2D-SIM

Figure 4 presents an investigation into the feature space of 2D-SIM, illustrating its enhanced capability in learning discriminative features for various human joints in comparison to a baseline model. This analysis was conducted by selecting a subset of videos from the Toyota-Smarthome dataset. The methodology involved computing the average distance of features between tokens corresponding to human joints across different layers within the video transformer. We find that towards the later layers of the video transformer, 2D-SIM is able to refine the representations to better disambiguate between the various human joints.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1, 4

[2] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. MARS: Motion-Augmented RGB Stream for Action Recognition. In *CVPR*, 2019. 1

[3] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Int. Conf. Comput. Vis.*, 2019. 1

[4] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *European Conference on Computer Vision*, pages 72–90. Springer, 2020. 1, 2

[5] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

[6] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2959–2968, 2022. 2

[7] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. *Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018. 1

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[9] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[10] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2021. 2

[11] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*

2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 12493–12506, 2021. 2

[12] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[13] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1

[14] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129:2264 – 2287, 2019. 1

[15] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022. 2, 4