From Video Classification to Action Detection: Foundation vs. Task-Specific Models

Gonçalo Mesquita [*1] **Ana Rita Cóias** [*1] **Artur Dubrawski** [2] **Alexandre Bernardino** [1]

Abstract

Real-time action detection demands fine-grained supervision, yet most skeleton based datasets only provide video-level annotations, due to the high cost, subjectivity, and time-consuming nature of frame-level labeling. To bridge this gap, we propose a pipeline that transforms video-level annotations into frame-level pseudo-labels via saliency maps. This approach significantly reduces the need for manual labeling while enabling frame-level action detection. We evaluate our method using both structured foundation models and task-specific architectures for action recognition (daily activities and rehabilitation) across four diverse datasets: SERE, Toronto Rehab, UTK and MMAct. These results highlight the generalization potential across users of the foundation models trained on structured time-series data, offering an efficient route from video-level labels to fine-grained motion analysis.

1. Introduction

Real-time action detection is a key component in numerous real-world applications, including human-computer interaction (Mitra & Acharya, 2007), robotics (Olatunji, 2018), rehabilitation and physical therapy monitoring (Cóias et al., 2022), sports performance feedback (Bialkowski et al., 2014), and intelligent surveillance systems (Kulbacki et al., 2023). These domains require not only the recognition of human actions, but also the immediate feedback of ongoing behaviors, which in turn requires frame-level annotations (Shou et al., 2016). However, most large-scale human action datasets, such as HMDB-51 (Kuehne et al., 2011), Kinetics-700 (Tölgyessy et al., 2021) and NTU RGB+D (Shahroudy et al., 2016; Liu et al., 2020) provide only video-level labels, limiting model development to offline classification and making them unsuitable for tasks that require temporal precision. Obtaining frame-level ground truth is notoriously difficult. Frame-level annotations are labor intensive, time-consuming (Sigurdsson et al., 2016), and susceptible to interannotator bias (Sigurdsson et al., 2016). Disagreements about action onset and offset, or divergent interpretations of ambiguous motions, introduce noise that can hinder training and generalization, especially in safety-critical or interactive systems where precision is important.

In real-world applications where timely and accurate feedback is essential, structured representations of human motion play a critical role in enabling robust and scalable perception systems. Skeleton data, which captures the spatial and temporal dynamics of body joints, provides a semantically rich, yet compact format for modeling human actions. In light of these challenges and using this structured data as input, we propose a novel framework that aims to improve pseudo-labels to the frame-level on action recognition by addressing the following: (i) **Saliency-based pseudo-labelling for skeleton data:** We propose a two-stage pipeline that first performs video-level action classification and then generates frame-level pseudo-labels using saliency maps and the pseudo-label selection method. This method enables weakly supervised action detection by identifying the most important frames, without requiring dense annotation. (ii) **Foundation vs. task-specific model comparison:** : We provide the first empirical comparison between a time-series foundation model (moment) and specialized skeletal transformers (Action Transformer and SkateFormer) across video classification, pseudo-labeling, and frame-level tasks. (iii) **Cross-subject evaluation on four benchmark datasets:** We evaluate our models on four datasets: Toronto Rehab (Dolatabadi et al., 2017) and SERE (Cóias et al., 2025), which focus on rehabilitation exercises, and UTK (Xia et al., 2012) and MMAct (Kong et al., 2019), which target daily activity recognition. Each dataset is adapted to support both video-level and frame-level evaluation. To assess cross-person generalization, we apply a Leave-One-Subject-Out (LOSO) cross-validation protocol across all experiments.

[*]Equal contribution [1]Institute for Systems and Robotics, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal [2]Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Gonçalo Mesquita <goncalo.mesquita@tecnico.ulisboa.pt>, Ana Rita Cóias <ana.coias@tecnico.ulisboa.pt>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

2. Related Work

Recent work has explored how gradient-based techniques can enhance temporal understanding and interpretability in video action recognition. In rehabilitation contexts, saliency methods such as vanilla and integrated gradients have been used to detect subtle anomalies and interpret model decisions effectively (Cóias et al., 2022; Lee, 2024). These approaches are especially valuable in settings where precise frame-level annotations are unavailable. More broadly, gradient-based methods like Grad-CAM (Selvaraju et al., 2017) and recent extensions such as the To-a-T Spatio-Temporal Focus framework (Ke et al., 2022) have demonstrated success in highlighting informative regions in temporal sequences, though they are underexplored in the context of skeleton based motion data. Skeleton based systems offer a lightweight and interpretable representation for modeling human motion. Early methods relied on binary classifiers for stroke rehabilitation (Lee et al., 2019), while more recent approaches leverage webcam-based pose estimation combined with machine learning to detect compensatory behaviors with greater accuracy and generalizability (Cóias et al., 2023). Finally, the choice between foundation and task-specific models continues to shape advances in structured prediction. Foundation models like Moment (Goswami et al., 2024) offer strong generalization via large-scale pretraining across diverse time-series tasks. In contrast, specialized models such as Action Transformer (AcT) (Mazzia et al., 2022) and SkateFormer (SF) (Do & Kim, 2024) are explicitly designed for human motion analysis, incorporating skeletal priors and attention mechanisms that excel in domain-specific applications.

3. Method

3.1. Approach Pipeline Overview

Figure 1 shows the complete pipeline we adopt for real-time action detection in untrimmed videos. Consider the dataset $V = \{v^i\}_{i=1}^N$ and its associated video-level labels $y^i \in \{0,1\}^K$; the target $y_k^i = 1$ indicates that action k occurs somewhere in v^i, making the task binary multi-label when $K > 1$. We first extract body-pose keypoints with a state-of-the-art detector and denoise them in a preprocessing step (box a). These features, together with the video-level labels, are used to fine-tune video classifiers (Model A in box b), of architectures AcT (Mazzia et al., 2022), SF (Do & Kim, 2024), Moment (Goswami et al., 2024), or an LSTM. Next, we apply a gradient-based saliency method (Simonyan et al., 2014) (Sundararajan et al., 2017) to Model A's predictions to obtain temporal saliency scores. The Pseudo-Label Selection method thresholds these scores to construct binary frame-level labels $z_t^i \in \{0,1\}^K$, setting $z_{t,k}^i = 1$ whenever the saliency for action k at frame t exceeds the predefined cut-off (box c). Finally, the same feature sequence f^i together with the derived pseudo-labels $\{z_t^i\}$ are used to train a multilayer perceptron (MLP) for frame-level action recognition (box d), whose outputs are evaluated against the ground truth annotations.

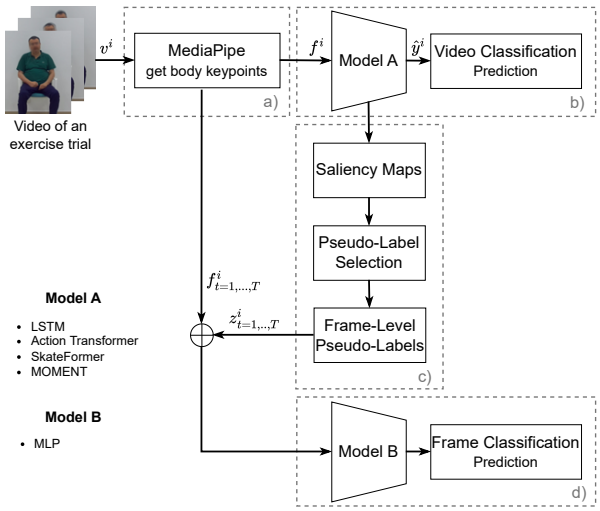

Figure 1. Approach Pipeline: (a) Body pose extraction and preprocessing. (b) Video-level classification using Model A. (c) For positive predictions, saliency maps are generated to create frame-level pseudo-labels. (d) An MLP (Model B) is trained on these pseudo-labels for frame-level assessment.

3.2. Structured Pose Representation and Preprocessing

To impose a structured inductive bias over body dynamics, we use 2D/3D skeletal keypoints as input features. For each dataset, we apply the appropriate pose extractor, MediaPipe for SERE (Cóias et al., 2023), Microsoft Kinect (Tölgyessy et al., 2021) for Toronto Rehab and UTK (Xia et al., 2012), and OpenPose (Cao et al., 2019; Simon et al., 2017; Cao et al., 2017; Wei et al., 2016) for MMAct (Kong et al., 2019) ensuring that each frame is represented by a set of joint coordinates. To reduce inter-subject variability, we normalize joint positions by subtracting their initial positions within a trial, capturing displacement vectors instead of absolute locations. Additionally, we apply a temporal smoothing filter (a five-frame moving average) to suppress sensor noise and increase signal coherence over time.

3.3. Video-Level Action Classification: Foundation vs. Task-Specific Models

To assess the role of model architecture and pretraining strategy in structured prediction, we compare both task-specific and foundation models in their ability to classify human motion (in rehabilitation settings and daily activities) using only video-level annotations. We include the AcT, a transformer-based model that leverages full self-attention over short sequences of body keypoints and is pre-trained

on MPOSE2021, a pose-centric dataset. SF (Do & Kim, 2024) introduces an inductive bias by grouping joints into semantically meaningful skeletal-temporal partitions, applying attention within and across these groups to better capture joint interactions. We also evaluate Moment (Goswami et al., 2024), a foundation model pre-trained across diverse time-series domains and tasks such as classification and anomaly detection. This model serves as a representative of broadly pre-trained structured models and is expected to exhibit strong adaptability to novel domains. A LSTM baseline was also trained from scratch, providing a minimal inductive bias comparator without any pre-trained knowledge. All models transformer based models are fine-tuned solely with video-level labels. By comparing their downstream performance and the quality of the saliency-derived pseudo-labels, we isolate how structural priors and pretraining strategies affect the learning.

3.4. Frame-Level Pseudo-Labels Generation

After training the video-level classifier (Model A), we use gradient based saliency methods, Vanilla Gradient (VG) (Simonyan et al., 2014) and Integrated Gradients (IG) (Sundararajan et al., 2017), to identify frames most influential to the model's predictions. These techniques assign importance scores to input features, revealing which frames are most relevant for classification. While Vanilla Gradient measures local sensitivity, Integrated Gradients provide more stable attributions by integrating over a baseline-to-input path.

3.5. Pseudo-label Selection Method

From the saliency maps, we aggregate gradients frame by frame and apply min-max normalization to scale the results to a range of $[0, 1]$, yielding a pseudo-score s_t^i for each frame t. We distinguish the frames of a motion labeled negative from the frames of a motion labeled positive by thresholding frames pseudo-scores. We explore a technique we call single threshold. This approach requires only one threshold τ. Using a threshold, τ, each frame is assigned with a pseudo-label, z_t^i,

$$z_t^i = \begin{cases} 0, \text{ if } \hat{y}^i = 0 \\ \mathbb{I}(s_t^i > \tau), \text{ if } \hat{y}^i = 1 \end{cases} \tag{1}$$

where $\hat{y}^i$ represents the predicted class for video i, and $\mathbb{I}$ is the indicator function. For normal motion video trial ($\hat{y}^i = 0$), all frames are assigned with a pseudo-label $z_t^i = 0$. For videos with positive label motions ($\hat{y}^i = 1$), each frame's pseudo-score s_t^i is compared against the threshold τ. If $s_t^i > \tau$, the indicator function assigns a frame pseudo-label $z_t^i = 1$ and $z_t^i = 0$, otherwise. The threshold τ used to convert the pseudo-saliency scores into binary frame-level labels was selected subjectively for each dataset based on inspection of the pseudo-score distributions. While τ differs across datasets due to variations in label density, action granularity and skeleton algorithm, it is held constant across all models within a given dataset. This ensures fair model comparisons under consistent pseudo-label conditions.

3.6. Frame-Level Compensation Classifier

In Model B, we implement a Multilayer Perceptron (MLP) to evaluate the effectiveness of these pseudo-labels. For comparison, we also train the same model using ground-truth (GT) frame-level annotations, providing an upper-bound performance that reflects the best achievable results under our evaluation protocol.

4. Datasets

We evaluate our method on four datasets spanning rehabilitation and general action recognition. Each dataset is adapted to support both video-level and frame-level evaluation. For both classification tasks, we adopt a LOSO) cross-validation protocol to assess model generalization across individuals.

Toronto Rehab Upper-Limb Dataset (Dolatabadi et al., 2017): This dataset includes 19 participants performing upper-limb exercises. Data were recorded with a Microsoft Kinect v2 sensor at 30 FPS, capturing 3D joint trajectories. Expert frame-level annotations identify four compensation types. Since only frame-level labels are provided, we created video-level labels by marking a video as compensatory ($\hat{y}^i = 1$) if any frame in the sequence is annotated as compensatory, and non-compensatory otherwise ($\hat{y}^i = 0$).

MMAct Dataset (Kong et al., 2019): Originally designed for multi-class action recognition (daily activities), MMAct contains over 36,000 labeled video clips from 20 participants across 37 action classes, captured across four environments using multi-view cameras and multiple sensor modalities. For our binary detection task, we concatenate three action clips into a single video. The video-level label is set to positive if the target action is present in any of the three clips. The frame-level labels, initially provided for each clip, are preserved in each frame of the clip.

SERE Dataset (Cóias et al., 2025): This dataset consists of 1,260 short video clips recorded at 30 FPS from 18 post-stroke patients performing five rehabilitation tasks. Video and frame-level annotations of compensatory behavior are provided by expert clinicians, with binary labels indicating the presence (1) or absence (0) of compensation. The dataset captures realistic variability in clinical movement patterns, making it a valuable testbed for generalization.

UTK (Xia et al., 2012): This dataset is a benchmark dataset designed for human action recognition using 3D skeletal data. It was collected using a Microsoft Kinect sensor and

includes 10 different action classes (daily activities) performed by 10 subjects from varying viewpoints. For our binary detection task, we concatenate three action clips into a single video. The video-level label is set to positive if the target action is present in any of the three clips. The frame-level labels, initially provided for each clip, are preserved in each frame of the clip.

5. Results and Experiments

5.1. Video-level Classification

Table 1 presents the video-level classification performance across four datasets using four different models, evaluated under a LOSO cross-validation protocol. Moment consistently achieved the highest AUC values, excelling on the SERE dataset, UTK dataset and MMAct. Therefore being able to create the best embeddings to detect human motion across different people. SF recorded the best performance on Toronto, highlighting its task-specific effectiveness. In contrast, the baseline LSTM model, which lacks pretraining and structural priors, trailed behind the transformer-based models. Interestingly, performance variances are lower on MMAct and UTK, likely because Toronto and SERE are rehabilitation datasets, where patient-specific motion patterns make generalization across subjects more difficult.

Method	**Toronto**$_{\text{N=19}}$	**MMAct**$_{\text{N=20}}$	**SERE**$_{\text{N=18}}$	**UTK**$_{\text{N=10}}$
LSTM	0.75 ± 0.05	0.93 ± 0.01	0.58 ± 0.06	0.97 ± 0.01
AcT	0.71 ± 0.07	0.96 ± 0.01	0.67 ± 0.05	0.92 ± 0.03
SF	$\mathbf{0.76 \pm 0.05}$	0.96 ± 0.01	0.65 ± 0.05	0.96 ± 0.02
Moment	0.72 ± 0.04	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.73 \pm 0.05}$	$\mathbf{0.98 \pm 0.01}$

Table 1. Video-level classification results (AUC) under LOSO cross-validation across four datasets. Values are reported as mean $\pm$ standard error. Sample size N indicates the number of subjects.

5.2. Frame-level Pseudo-label Classification

Moving to the more challenging frame-level task, Table 2 shows that IG consistently outperform VG across all models and datasets. This confirms that saliency-based attribution significantly improves the ability to identify discriminative frames from video-level supervision. Among all models, AcT consistently produces the most informative saliency maps. The gains from IG are especially pronounced for AcT, for example, on the SERE dataset, its AUC rises from 0.53 (VG) to 0.72 (IG), marking the highest jump among all models and surpassing ground-truth performance. In contrast, Moment, despite its strong video-level classification performance, yields saliency maps that are generally less aligned with frame-level relevance. Its frame-level AUCs improve moderately with IG but remain behind AcT and SF on most datasets. This suggests that while Moment captures broad motion representations that generalize across subjects, its internal attention may be less focused on temporally discriminative regions. SF also benefits from IG, particularly on MMAct, where it achieves the highest AUC (0.86), but it trails behind AcT on rehabilitation datasets like Toronto and UTK. Across all datasets, our pseudo-labeling pipeline consistently narrows the AUC gap of the best performer to ground truth labels by only 4–12 points, offering significant annotation cost savings with minimal performance loss. Specifically, the Toronto dataset showed a 10-point drop from 0.80 to 0.70, MMAct had a 5-point decline from 0.91 to 0.86, and UTK exhibited the largest reduction, with a 12 point difference from 0.91 to 0.79. Interestingly, on the SERE dataset, pseudo-labeling with AcT_{IG} outperformed the ground truth, achieving an AUC of 0.72 compared to 0.69 with expert annotations, suggesting that the manually labeled data may be affected by human bias.

Method	**Toronto**$_{\text{N=19}}$	**MMAct**$_{\text{N=20}}$	**SERE**$_{\text{N=18}}$	**UTK**$_{\text{N=10}}$
LSTM_{VG}	0.43 ± 0.03	0.83 ± 0.02	0.64 ± 0.02	0.52 ± 0.05
LSTM_{IG}	0.65 ± 0.03	0.82 ± 0.02	0.64 ± 0.03	0.76 ± 0.03
AcT_{VG}	0.45 ± 0.02	0.85 ± 0.02	0.53 ± 0.04	0.48 ± 0.06
AcT_{IG}	$\mathbf{0.70 \pm 0.03}$	0.85 ± 0.02	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.79 \pm 0.03}$
SF_{VG}	0.63 ± 0.03	0.84 ± 0.02	0.58 ± 0.02	0.52 ± 0.04
SF_{IG}	0.63 ± 0.02	$\mathbf{0.86 \pm 0.02}$	0.66 ± 0.03	0.78 ± 0.03
$\text{Moment}_{\text{VG}}$	0.47 ± 0.02	0.85 ± 0.02	0.51 ± 0.02	0.49 ± 0.05
$\text{Moment}_{\text{IG}}$	0.61 ± 0.03	0.84 ± 0.02	0.69 ± 0.02	0.78 ± 0.03
GT	0.80 ± 0.03	0.91 ± 0.02	0.69 ± 0.03	0.91 ± 0.02

Table 2. Frame-level classification results (AUC, mean $\pm$ standard error) from LOSO cross-validation using Vanilla Gradients (VG) and Integrated Gradients (IG) for pseudo-labeling across four datasets. Sample size N indicates the number of subjects.

6. Conclusion and Limitations

This work demonstrates that gradient-based saliency methods, particularly IG, can effectively bridge the gap between video-level labels and strong frame-level annotations. While our approach shows strong potential, it has two main limitations. First, the reliance on a manual threshold, which harms reproducibility. Second, although Moment performed well in video classification, it was less temporally focused compared to the task-specific, making its pseudo-labels less informative in some contexts.

Overall, our findings suggest that saliency-guided pseudo-labeling, when combined with either a pre-trained foundation model or a task-specific skeletal transformer, provides a practical and effective strategy for fine-grained action recognition without the need for dense annotations.

Acknowledgements

The authors would like to thank their colleagues from the Robot Vision Laboratory (VisLab), Laboratory of Robotics and Engineering Systems (LARSyS), Instituto Superior Técnico, and Mononito Goswami of the CMU Auton Lab, Carnegie Mellon University, for their insightful discussions, technical support, and valuable feedback throughout this research.

This work was supported by the Portuguese Foundation for Science and Technology - FCT by LARSyS FCT funding (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020), FCT HAVATAR Project (DOI: 10.54499/PTDC/EEI-ROB/1155/2020), by FCT under CMU Portugal, by the Portuguese Recovery and Resilience Plan (RRP), through project number 62, Center for Responsible AI, by NSF awards 2427948 and 2406231, and by DARPA award HR00112420329. Ana Rita Cóias is supported by the FCT doctoral grant [SFRH/BD/05239/2021].

Impact Statement

This paper presents work that aims to advance the field of Weakly Supervised Machine Learning techniques and explore the application of MOMENT, the time-series foundation model, for action recognition and motion quality assessment. This contribution can enhance the development of fitness and rehabilitation support tools. There are many other potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bialkowski, A., Lucey, P., Carr, P., and Matthews, I. Real-time sports tracking and analysis: A comprehensive survey. *IEEE Signal Processing Magazine*, 31(4):118–133, 2014.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Cóias, A. R., Lee, M. H., and Bernardino, A. A low-cost virtual coach for 2D video-based compensation assessment of upper extremity rehabilitation exercises. *Journal of NeuroEngineering and Rehabilitation*, 19 (1):83, July 2022. ISSN 1743-0003. doi: 10.1186/s12984-022-01053-z. URL https://doi.org/10.1186/s12984-022-01053-z.

Cóias, A. R., Lee, M. H., Bernardino, A., and Smailagic, A. Skeleton Tracking Solutions for a Low-Cost Stroke Rehabilitation Support System. In *2023 International Conference on Rehabilitation Robotics (ICORR)*, pp. 1–6, September 2023. doi: 10.1109/ICORR58425.2023.10304749. URL https://ieeexplore.ieee.org/document/10304749. ISSN: 1945-7901.

Cóias, A. R., Lee, M. H., Bernardino, A., Smailagic, A., Mateus, M., Fernandes, D., and Trapola, S. Learning frame-level classifiers for video-based real-time assessment of stroke rehabilitation exercises from weakly annotated datasets. *TechRxiv*, January 2025. doi: 10.36227/techrxiv.173834586.68270298/v1.

Do, J. and Kim, M. SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition, July 2024. URL http://arxiv.org/abs/2403.09508. arXiv:2403.09508 [cs].

Dolatabadi, E., Zhi, Y., Ye, B., Coahran, M., Lupinacci, G., Mihailidis, A., Wang, R., and Taati, B. *The toronto rehab stroke pose dataset to detect compensation during stroke rehabilitation therapy*. May 2017. doi: 10.1145/3154862.3154925. Pages: 381.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

Ke, L., Peng, K.-C., and Lyu, S. Towards to-a-t spatio-temporal focus for skeleton-based action recognition. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://arxiv.org/abs/2202.02314.

Kong, Q., Wu, Z., Deng, Z., Klinkigt, M., Tong, B., and Murakami, T. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8657–8666, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00875. URL https://ieeexplore.ieee.org/document/9009579/.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. HMDB: A large video database for human motion recognition. In *International Conference on Computer Vision*, pp. 2556–2563, 2011.

Kulbacki, M., Segen, J., Chaczko, Z., Rozenblit, J. W., Kulbacki, M., Klempous, R., and Wojciechowski, K. Intelligent Video Analytics for Human Action Recognition: The State of Knowledge. *Sensors*, 23(9):4258,

January 2023. ISSN 1424-8220. doi: 10.3390/s23094258. URL https://www.mdpi.com/1424-8220/23/9/4258. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

Lee, M. H. Towards Gradient-based Time-Series Explanations through a SpatioTemporal Attention Network, May 2024. URL http://arxiv.org/abs/2405.17444. arXiv:2405.17444 [cs].

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and Badia, S. B. i. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pp. 218–228, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. doi: 10.1145/3301275.3302273. URL https://dl.acm.org/doi/10.1145/3301275.3302273.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, October 2020. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2019.2916873. URL http://arxiv.org/abs/1905.04757. arXiv:1905.04757 [cs].

Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., and Chiaberge, M. Action Transformer: A Self-Attention Model for Short-Time Pose-Based Human Action Recognition. *Pattern Recognition*, 124:108487, April 2022. ISSN 00313203. doi: 10.1016/j.patcog.2021.108487. URL http://arxiv.org/abs/2107.00606. arXiv:2107.00606 [cs].

Mitra, S. and Acharya, T. Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, May 2007. ISSN 1558-2442. doi: 10.1109/TSMCC.2007.893280. URL https://ieeexplore.ieee.org/document/4154947/.

Olatunji, I. E. Human Activity Recognition for Mobile Robot. *Journal of Physics: Conference Series*, 1069:012148, August 2018. ISSN 1742-6588, 1742-6596. doi: 10.1088/1742-6596/1069/1/012148. URL https://iopscience.iop.org/article/10.1088/1742-6596/1069/1/012148.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis, April 2016. URL http://arxiv.org/abs/1604.02808. arXiv:1604.02808 [cs].

Shou, Z., Wang, D., and Chang, S.-F. Temporal action localization in untrimmed videos via multi-stage CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1049–1058, 2016.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding, July 2016. URL http://arxiv.org/abs/1604.01753. arXiv:1604.01753 [cs].

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. URL http://arxiv.org/abs/1312.6034. arXiv:1312.6034 [cs].

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic Attribution for Deep Networks, June 2017. URL http://arxiv.org/abs/1703.01365. arXiv:1703.01365 [cs].

Tölgyessy, M., Dekan, M., and Chovanec, Skeleton Tracking Accuracy and Precision Evaluation of Kinect V1, Kinect V2, and the Azure Kinect. *Applied Sciences*, 11(12):5756, January 2021. ISSN 2076-3417. doi: 10.3390/app11125756. URL https://www.mdpi.com/2076-3417/11/12/5756. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. Convolutional pose machines. In *CVPR*, 2016.

Xia, L., Chen, C.-C., and Aggarwal, J. K. View invariant human action recognition using histograms of 3D joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, Providence, RI, USA, June 2012. IEEE. ISBN 978-1-4673-1612-5 978-1-4673-1611-8 978-1-4673-1610-1. doi: 10.1109/CVPRW.2012.6239233. URL http://ieeexplore.ieee.org/document/6239233/.