Pose-Anchored Optical Flow for Low-Latency Human Action Anticipation in Human-Robot Teaming

Lewis de Zoete Grundy[1], Chris McCarthy[1], and Christopher Fluke[1]

***Abstract*— Human-robot interaction (HRI) requires robots to interpret human actions early in their execution in order to respond safely, efficiently, and naturally. However, many existing approaches to human action recognition rely either on sparse skeletal representations, which lack fine-grained motion cues, or dense optical flow, which can be computationally expensive for low-latency perception pipelines. In this paper, we propose PoseOFF, a pose-anchored optical flow representation that captures local motion information around human joints to support earlier human intent understanding. By conditioning motion feature extraction on human pose, PoseOFF encodes localised motion dynamics at semantically meaningful body locations, forming a structured motion representation that is explicitly aligned with human kinematics. We evaluate PoseOFF across multiple benchmark datasets and backbone architectures for action anticipation, demonstrating consistent improvements in recognition accuracy, particularly at early observation ratios. Our results show that PoseOFF enables models to achieve comparable or improved performance while observing less of the action sequence, highlighting its effectiveness for early prediction. Importantly, these gains are achieved without requiring full-frame motion processing, making the approach practical for real-time and resource-constrained settings. These findings suggest that pose-centred motion representations such as PoseOFF can enhance the ability of interactive robot systems to infer human actions earlier, supporting more responsive and anticipatory behaviour in human-robot interaction scenarios.**

I. Introduction

Effective human-robot interaction (HRI) relies on a robot's ability to interpret human behaviour as it unfolds, rather than after it has completed. In collaborative environments, robots must anticipate human actions to coordinate movements, avoid interference, and ensure safety in shared workspaces [1], [2]. In assistive and social contexts, early recognition of user intent enables timely support and more natural interaction, supporting proactive assistance and fluid turn-taking. Across these domains, the ability to infer human intent from partial observations is central to responsive and human-aware robotic systems.

Despite significant progress in human action recognition, early action anticipation remains a challenging problem. Many approaches rely on skeletal representations derived from pose estimation, which provide a compact and semantically meaningful description of human body configuration [3], [4]. These representations are well suited to real-time systems due to their efficiency and robustness to background variation. However, skeleton-based methods inherently discard fine-grained motion information, limiting their ability to distinguish between actions with similar pose configurations but different movement dynamics, particularly in the early stages of an action.

Fig. 1: PoseOFF captures localised motion patterns (e.g., rotation, divergence) around key joints that are not represented in skeletal pose alone.

Conversely, dense motion representations such as optical flow capture rich spatiotemporal information and have been widely used to improve action recognition performance [5]. More recent work has explored transformer-based architectures for action anticipation, achieving strong performance through large-scale spatiotemporal modelling [6]. However, these approaches often come with increased computational cost, making them less suitable for low-latency perception pipelines required in interactive robotic systems.

This trade-off between efficiency and motion fidelity presents a key challenge for deploying action anticipation methods in HRI settings. A promising direction is to focus on local, body-centred motion cues that retain informative dynamics while avoiding the cost of full-frame processing. By anchoring motion representations to human pose keypoints, it is possible to capture the most relevant movement patterns associated with human actions, while maintaining a compact and interpretable representation. Such approaches align well with the requirements of HRI systems, where perception modules must operate efficiently and provide timely insights into human behaviour.

In HRI settings, earlier recognition is often more valuable than marginal gains in full-sequence classification accuracy

[1]School of Science, Computing and Emerging Technologies, Swinburne University of Technology, John St, Hawthorn VIC 3122, Australia {ldezoetegrundy,cdmccarthy,cfluke}@swin.edu.au

because it directly affects response timing, coordination quality, and safety margins. Prior work in HRI has shown that the timing and predictability of robot responses play a critical role in interaction fluency and human trust [7]. A method that can reliably infer human actions from partial observations—even with slightly lower peak accuracy—may therefore be more useful in practice than one that performs best only after the action has fully unfolded.

While prior work has explored combining pose and motion information, these approaches typically rely on either parallel processing streams or global motion representations, without explicitly structuring motion features around human kinematics. As a result, motion information is often treated as a dense signal rather than a behaviourally grounded representation aligned with the human body. This limits both interpretability and efficiency, particularly in early action anticipation, where informative motion cues are localised and temporally sparse.

Motivated by this, we propose PoseOFF, a pose-conditioned motion representation that encodes localised optical flow features at human joint locations that captures local motion information around human joints to support early action anticipation. PoseOFF samples optical flow in the vicinity of estimated keypoints, producing a compact description of motion that complements skeleton-based representations. Importantly, this representation can be integrated into existing architectures with minimal modification, providing a lightweight mechanism for enhancing early human intent understanding in interactive robot systems.

We evaluate PoseOFF across multiple benchmark datasets and backbone models for action anticipation, demonstrating consistent improvements in performance at early observation ratios. Our results show that incorporating local motion cues enables models to achieve comparable or improved recognition accuracy while observing less of the action sequence. These findings highlight the potential of pose-centred motion representations as a practical enhancement for low-latency perception pipelines in HRI applications.

In summary, this paper makes the following contributions:

- We propose PoseOFF, a novel pose-conditioned motion representation that anchors optical flow to human joints, introducing a structured, body-aligned alternative to dense or parallel motion representations.
- We show that this representation enables earlier and more discriminative action anticipation, achieving comparable accuracy with reduced observation.
- We demonstrate the value of targeted, low-latency motion representations for HRI, supporting earlier and more responsive human–robot interaction.

This paper is structured as follows: Section II reviews related work. Section III presents the PoseOFF representation. Section IV reports experimental results, including action anticipation, per-class analysis, and performance considerations. Section V concludes with discussion and future directions.

II. Related work

A. Human Action Anticipation for Interactive Systems

Human action anticipation aims to recognise actions from partial observations, enabling early inference of human intent. This capability is particularly important in interactive robot systems, where timely interpretation of human behaviour supports safe collaboration and responsive interaction [7], [1].

Early approaches relied on probabilistic models and handcrafted features [8], while later work introduced recurrent architectures to encourage prediction from partial sequences [9]. More recently, transformer-based models such as AVT [6] have demonstrated strong performance by modelling long-range temporal dependencies, with multimodal approaches further incorporating complementary cues [10]. However, these methods typically rely on dense spatiotemporal representations and large models, limiting their suitability for real-time deployment in interactive robotic systems.

B. Skeleton-based Human Action Recognition

Skeleton-based representations provide a compact and semantically meaningful description of human motion through joint trajectories, and are widely used due to their efficiency and robustness to background variation [3], [4]. Graph-based approaches such as ST-GCN [3] model spatial and temporal relationships between joints, while extensions such as 2s-AGCN [4] introduce adaptive graph structures and motion streams to improve representation capacity. More recent methods, including PoseC3D [11], explore alternative formulations using spatiotemporal convolutions over pose sequences.

These approaches achieve strong performance while maintaining relatively low computational cost, making them well suited to real-time and embedded systems, including many human-robot interaction scenarios. However, skeleton-based representations abstract away fine-grained motion information present in the underlying visual signal. While joint trajectories capture coarse movement patterns, they do not explicitly encode local motion dynamics such as subtle limb movements, object interactions, or inter-frame motion cues.

This limitation becomes particularly significant in early action anticipation, where partial observations may not yet exhibit distinctive pose configurations. In such cases, informative motion cues may be present even when pose alone is ambiguous, suggesting that augmenting skeleton-based representations with additional motion information is important for improving early prediction performance.

C. Motion-based and Multimodal Approaches

Motion-based representations, particularly those based on optical flow, capture dynamic information that complements pose and appearance features. Two-stream architectures [5] and more recent video models [12] demonstrate the importance of motion cues for action understanding [10]. However, these approaches typically model motion as a dense, global signal, leading to substantial computational cost and redundancy. While multimodal fusion improves

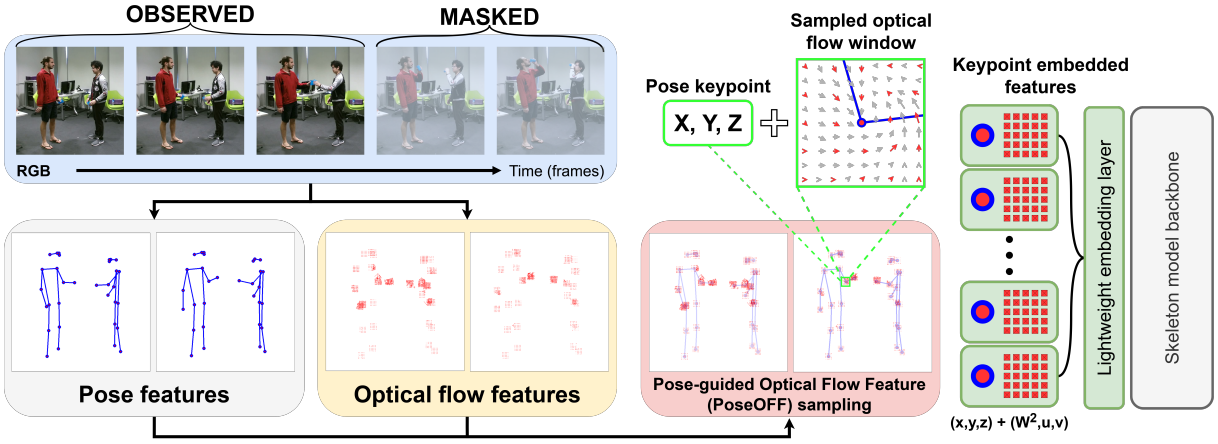


Fig. 2: PoseOFF representation construction. From an input RGB video, human pose keypoints and optical flow fields are extracted. Motion features are computed at each joint location by encoding local optical flow dynamics in the vicinity of the keypoint. These pose-conditioned motion features are combined with joint coordinates to form a structured representation aligned with human kinematics, capturing localised motion information for each body part. Motion feature windows are concatenated with keypoint information without alteration to skeleton joint relations. A lightweight embedding layer allows the PoseOFF data to be fed into an existing skeleton-based action recognition models without modification to model backbone.

recognition performance, it does not explicitly structure motion representations around human kinematics. As a result, motion information is often encoded in a manner that is both computationally expensive and weakly aligned with human body dynamics, limiting its effectiveness for efficient and early action understanding.

Joint-centred motion representations, such as JOLO-GCN [13] and H-MoRe [14], have demonstrated that visual motion surrounding human body keypoints provides useful learnable information. Critically, however, these methods treat joint-centred motion as a separate, complementary input stream to skeleton keypoints rather than integrating motion features within the skeleton-based learning process itself. Moreover, both JOLO-GCN and H-MoRe are designed and evaluated specifically for action recognition, and have not been extended to action anticipation, where predictions must be made from partial, early observations of an unfolding action.

Across these approaches, a consistent trade-off emerges. Skeleton-based methods are efficient but lack fine-grained motion cues, while motion-based and multimodal methods capture richer dynamics at significantly higher computational cost. In interactive robotic systems, where perception must be both timely and resource-efficient, this trade-off becomes critical [7], [2].

This reveals a clear gap in the literature. No existing work has sought to integrate joint-centred motion features directly within the skeleton-based action recognition framework and apply this to the task of action anticipation. Such integration would allow motion cues to interact with structural keypoint representations during learning, rather than being processed as a separate parallel signal, potentially yielding representations that are simultaneously more informative, better aligned with human body dynamics, and capable of supporting earlier and more accurate action predictions.

III. Method

We propose PoseOFF, a pose-conditioned motion representation designed to encode localised motion dynamics aligned with human body structure. Rather than modelling motion as a dense global signal or fusing it with pose at the feature level, PoseOFF explicitly conditions motion representation on human pose, capturing motion cues at semantically meaningful body locations.

This design is motivated by the need for efficient and early action understanding in interactive robotic systems, where informative motion cues are often localised and temporally sparse. By structuring motion features around human kinematics, PoseOFF provides a compact and behaviourally grounded representation that supports action anticipation while avoiding the computational overhead of dense motion modelling.

By embedding early motion cues as learnable features in skeleton-based action recognition architectures, we aim to show such features can support state-of-the-art models achieving comparable accuracy with fewer observed frames. Moreover, we seek to achieve higher action anticipation ability with minimal additional computational overheads, thus retaining the feasibility of the methods applied in real-time settings on standard computing hardware.

A. PoseOFF Representation

PoseOFF encodes motion information by conditioning optical flow features on human pose. For each detected skeleton, motion features are extracted at the spatial locations of body joints, producing a representation that explicitly aligns motion dynamics with human kinematics.

Unlike conventional optical flow representations, which model motion densely across the entire frame, PoseOFF focuses on local regions associated with human movement. This results in a structured motion representation that cap-

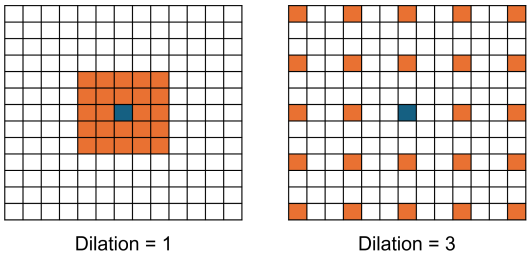


Fig. 3: Example of 5×5 optical flow sampling windows with dilation 1 (left) and 3 (right). The centre cell corresponds to the pose keypoint location..

tures behaviourally relevant dynamics while reducing redundancy.

The input to the PoseOFF representation is an RGB video consisting of T frames. For each frame I_t, where $t \in [1, T]$, up to M human skeletons are extracted, each defined by V keypoints. Each keypoint is represented by $C_{\text{pose}} = 3$ channels, corresponding to either (x, y, z) coordinates or (x, y, α), where α denotes pose estimation confidence.

To incorporate motion information, optical flow is computed between consecutive frames I_t and I_{t+1}, yielding $T-1$ flow fields with (u, v) vectors describing pixel-wise motion. Rather than representing motion densely across the entire frame, PoseOFF constructs motion features conditioned on the spatial configuration of the skeleton.

Specifically, for each keypoint location (x_t, y_t), a local neighbourhood of optical flow vectors is extracted, forming an $N \times N$ region of motion centred at the joint. This produces a local motion feature of shape $(N, N, 2)$, capturing the motion dynamics in the vicinity of each body part. A dilation factor may be applied to control the spatial extent of this region, allowing the representation to capture motion at different scales (see Figure 3).

Each local motion region is flattened to produce $C_{\text{flow}} = 2N^2$ channels, which are concatenated with the corresponding pose features. This results in a unified, pose-conditioned representation with $C = C_{\text{pose}} + C_{\text{flow}} = 2N^2 + 3$ channels per keypoint.

The final PoseOFF representation for a sequence is:

$$(T, M, V, C)$$

where motion features are explicitly aligned with joint locations, yielding a structured encoding of local motion dynamics.

B. Skeleton-based Action Recognition Models

To demonstrate the applicability and effectiveness of PoseOFF, we modify three state-of-the-art skeleton keypoint action recognition models to incorporate learnable early visual motion cues. The embedding layer of each model was modified to include a simple CNN designed to learn the features present in the windows of optical flow. The CNN consisted of two convolutional layers separated by activation layers, pooling and a linear layer. The output of these layers was concatenated with the original skeleton keypoints coordinates as additional channels in the skeleton graph. A diagram of the PoseOFF embedding layer added to each model is presented in Figure 4.

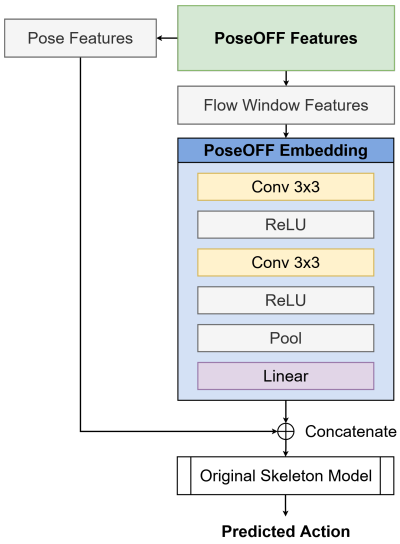


Fig. 4: PoseOFF embedding layer diagram - convolutional layers (Conv) learn from optical flow vector windows, then concatenate these embeddings with the original pose features (x,y,z positions).

The models evaluated were InfoGCN++ [15], MS-G3D [16], and ST-GCN++ [17], trained using hyperparametes specified by the original authors and settings as close to those reported in the respective papers.

The PoseOFF extraction method utilises human skeleton *joints* as locations in image space from which to extract windows of optical flow. For this reason, the results presented in this paper are models trained and evaluated using the *skeleton joint stream only*.

To evaluate early action recognition, we employ temporal masking to simulate partial observation of input sequences. Specifically, only the initial portion of each sequence is provided to the model, requiring classification based on incomplete action execution. Sequences are masked according to predefined *observation ratios*, which specify the proportion of frames retained as input. Figure 2 illustrates an example with a 60% observation ratio, where only the first part of the sequence is visible to the model. Partially observed sequences are padded to the required input length using the strategies defined in the original works: last frame duplication [15], sequence replay [16], and zero padding [17].

C. Implementation

All models were implemented in Python using the PyTorch framework [18], building on publicly available code released by the respective authors.

Optical flow was computed using the RAFT algorithm [19], with pretrained weights provided in the PyTorch ecosystem. Human pose keypoints were obtained directly from dataset annotations for NTU RGB+D 60 and NTU RGB+D 120. For the UCF101 dataset, pose keypoints were estimated using the YOLO-POSE Large model [20].

Training configurations were aligned with those reported in the original model implementations. For NTU RGB+D

datasets, hyperparameters were kept consistent with the respective baseline settings. For UCF101, where no standard configuration is defined, hyperparameters were selected to closely match those used for NTU RGB+D 60. In all cases, baseline and PoseOFF-augmented models were trained using identical settings to ensure fair comparison.

D. Datasets

The models were trained and evaluated on three different datasets; the NTU RGB+D 60 dataset [21], NTU RGB+D 120 dataset [22] and the UCF101 dataset [23].

The NTU RGB+D 60 and NTU RGB+D 120 datasets were captured using three Microsoft Kinect v2 sensors and provide ground-truth 3D human skeleton data. Each sample contains RGB videos, depth and infrared sequences, and 3D coordinates of 25 body joints. Standard train/test splits defined by the authors ensure fair and consistent evaluation. Samples are annotated with 60 and 120 action classes, respectively. Owing to their widespread use in skeleton-based action recognition research [15], [17], [16], these datasets were selected to establish baseline performance and to evaluate the impact of incorporating PoseOFF embeddings.

The UCF101 dataset consists of unconstrained, real-world videos depicting human actions. Skeleton keypoints and optical flow data were estimated using the methods described in Sec. III-C. The dataset includes three standard train/test splits. UCF101 was chosen to assess performance under more challenging, "in-the-wild" conditions, in contrast to the controlled capture settings of the NTU RGB+D datasets.

The RGB video data of the NTU RGB+D datasets has a resolution of 1920×1080 pixels, whereas the UCF101 dataset has a resolution of 320×240 pixels. Windows of optical flow for the PoseOFF embedding were sampled with a dilation factor of 3 for the NTU RGB+D datasets and a dilation factor of 1 for the UCF101 datasets. A higher dilation factor was chosen for the NTU RGB+D datasets given their higher video resolution. See Figure 3 an example of dilated window sampling.

IV. Experiments

A. Evaluation Metrics

For action classification, performance is measured using classification accuracy, defined as the percentage of samples for which the correct class label is predicted after observing the full action sequence. For action anticipation, we report classification accuracy at each observation ratio as well as the Area Under the Curve (AUC). The AUC summarizes performance across all evaluated observation ratios and is computed as:

$$\text{AUC} = \frac{\sum \text{Acc}_{\text{obs}}}{\text{N}_{\text{obs}}}$$

where Acc_{obs} denotes the accuracy at a given observation ratio, and N_{obs} is the number of evaluated observation ratios.

B. Action Classification

To evaluate action classification performance, both the base models and their PoseOFF-augmented variants (Figure 4) were trained and tested on full input sequences without temporal masking, simulating an offline action recognition setting.

Classification results are reported in Table I, with the best-performing model for each dataset and evaluation highlighted. Across all architectures, incorporating PoseOFF embeddings yields average accuracy improvements of 2.14%, 6.84%, and 11.22% on NTU RGB+D 60, NTU RGB+D 120, and UCF101, respectively, The improvements on UCF101 are particularly notable, as this dataset consists of RGB-only videos captured "in the wild," without depth or skeleton annotations. The skeletons extracted with YOLO-POSE can occasionally be misplaced or dropped due to the unconstrained nature of the videos in the dataset, and given the aggressively abstract nature of skeleton keypoints, misplaced keypoints can significantly impact action recognition performance. The addition of PoseOFF embeddings provides salient motion information and additional scene context, even when skeleton keypoints are misplaced.

C. Action Anticipation

To evaluate whether PoseOFF enables earlier action prediction, temporal masking was used to simulate partial observation of ongoing actions. MS-G3D and ST-GCN++ were trained on full sequences (100% observation) and evaluated at 10% observation ratio increments. InfoGCN++ natively performs frame-level prediction over a fixed 64-frame input, and is therefore evaluated directly without interval-based masking [15].

Results are shown in Fig. 5 for NTU RGB+D 60 (Cross-Subject) and UCF101. Performance is reported as classification accuracy at increasing observation ratios, with the baseline full-sequence accuracy indicated for reference. The point at which PoseOFF matches this baseline indicates how early equivalent recognition performance is achieved.

Across datasets, PoseOFF consistently enables models to reach baseline accuracy with fewer observed frames. On NTU RGB+D 60, ST-GCN++ and MS-G3D require only 80% of the sequence to match full-sequence baseline performance, while InfoGCN++ achieves equivalent accuracy after observing 50% of the input. Similar trends are observed on UCF101, where PoseOFF-augmented models match baseline performance after approximately half of the sequence.

These results demonstrate that PoseOFF improves early prediction capability, enabling comparable recognition performance under partial observation.

D. Per-Class Performance Analysis

To analyse where PoseOFF provides the greatest benefit, we compute per-class recall differences between baseline and PoseOFF models. Predictions are aggregated across models and confusion matrices are used to derive recall per class, providing a model-agnostic view of performance changes.

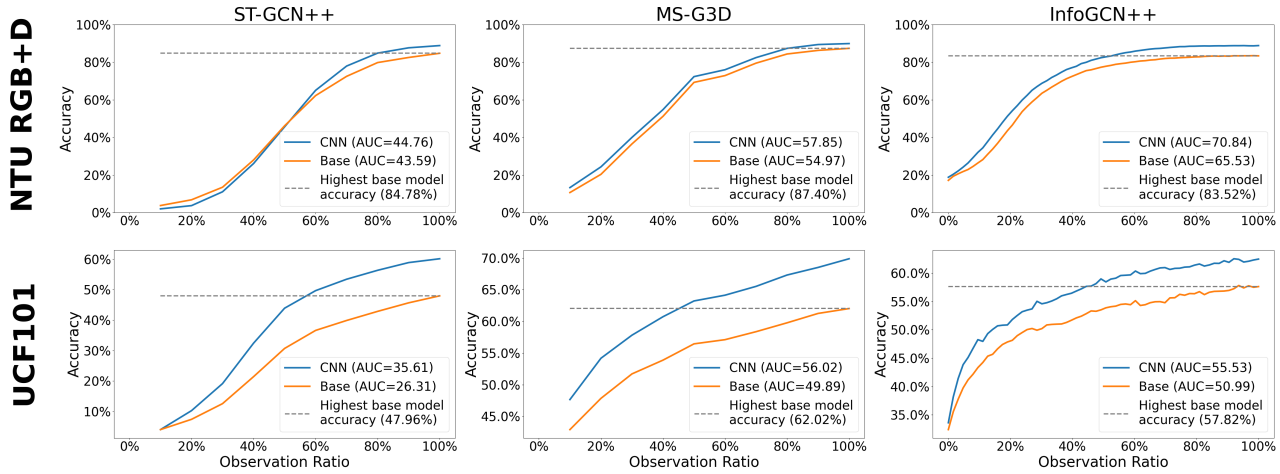


Fig. 5: Models trained using 100% observation ratios, evaluated at 10% observation ratio intervals on the NTU RGB+D 60 dataset, Cross Subject evaluation. The Area Under the Curve (AUC) is reported for each model in the sub figure labels.

Per-class recall differences for each dataset are visualised in Fig. 6a, Fig. 6b, and Fig. 6c. Across all datasets, PoseOFF yields consistent improvements across the majority of action classes. On NTU RGB+D 60, performance improves for 88.33% of classes, with a median recall gain of 2.12% ($p < 10^{-9}$). On NTU RGB+D 120, improvements are observed for 98.33% of classes, with a median gain of 5.98% ($p < 10^{-20}$). On UCF101, PoseOFF improves 80.20% of classes with a median gain of 7.65% ($p < 10^{-13}$). These results indicate that gains are broadly distributed rather than concentrated in a small subset of classes.

The distribution of improvements reveals clear trends across action categories. On NTU datasets (Fig. 6a, Fig. 6b), the largest gains are observed in daily actions involving fine-grained and localised motion, particularly hand-object interactions and self-directed actions (e.g., writing, typing, applying cream, cutting nails). These actions are characterised by subtle, local motion cues that are not fully captured by skeletal pose alone.

A similar pattern is observed on UCF101 (Fig. 6c), where improvements are strongest in categories involving structured body motion and object interaction, while gains are more variable for sports-related actions. In particular, sports classes—characterised by large-scale, full-body motion—exhibit both strong improvements and occasional performance drops, suggesting that global motion dynamics may not always be fully captured by localised representations.

To further interpret these trends, Fig. 7 presents qualitative examples of action classes where PoseOFF yields improved recall. These examples illustrate how localised motion cues captured by pose-anchored flow contribute to improved discrimination under challenging visual conditions.

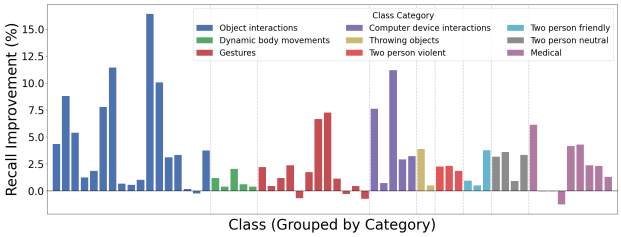
For instance, in *fall over*, limb occlusion limits the availability of optical flow for some joints; however, motion patches anchored to neighbouring visible limbs still capture informative dynamics, enabling correct classification. In *swing arms*, improvements arise from early detection of hand and elbow motion, which precedes significant changes in skeleton joint positions and would otherwise be weakly represented in pose-only features. In *cheers and drink*, PoseOFF captures rotational motion of the elbows, although the example also highlights the presence of noisy flow patches in visually cluttered scenes.

Across these examples, the motion fields surrounding key joints (e.g., elbows and knees) exhibit structured patterns, including curl in the flow field corresponding to limb rotation, and divergence where joints move toward or away from the camera. These local motion signatures are not explicitly encoded in skeletal representations but are captured by PoseOFF. Figure 1 provides two more examples (*pick-up* and *jump*) where PoseOFF yields improved recall.

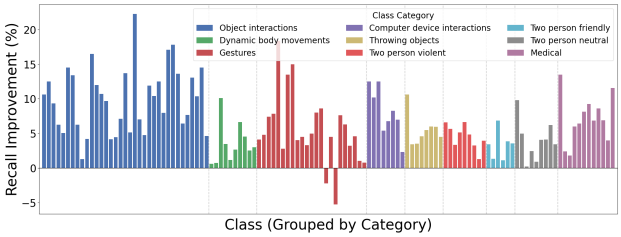
Conversely, performance decreases are limited to a small number of classes across all datasets. These typically correspond to actions dominated by global motion patterns or

TABLE I: Classification accuracy (%) for baseline and PoseOFF-augmented models across NTU RGB+D 60 (NTU), NTU RGB+D 120 (NTU120), and UCF101 datasets (joint stream only). Best results per model are in bold.

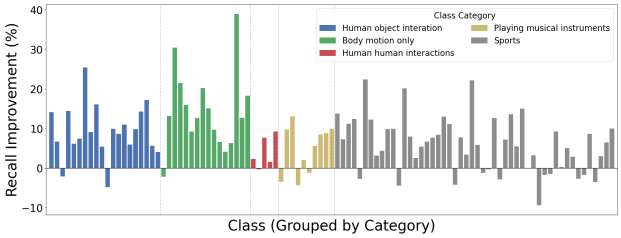
		NTU		NTU120		UCF101		
Model	**Variant**	**CS**	**CV**	**CSub**	**CSet**	**Split 1**	**Split 2**	**Split 3**
InfoGCN++	Base	83.52	91.38	75.92	77.61	59.21	57.82	58.90
	PoseOFF	**88.89**	**93.68**	**82.70**	**83.70**	**62.81**	**62.54**	**62.84**
MS-G3D	Base	87.40	93.43	76.24	78.75	61.39	62.02	63.45
	PoseOFF	**89.95**	**94.12**	**83.22**	**85.49**	**68.84**	**69.88**	**66.36**
ST-GCN++	Base	84.78	90.85	76.54	76.40	47.37	47.96	46.86
	PoseOFF	**88.88**	**93.70**	**82.36**	**85.00**	**58.27**	**60.13**	**58.61**



(a) NTU RGB+D 60



(b) NTU RGB+D 120



(c) UCF101

Fig. 6: Per-class recall improvement (PoseOFF vs baseline) for (a) NTU RGB+D 60, (b) NTU RGB+D 120, and (c) UCF101. Classes are grouped by action category.

complex multi-agent interactions, where localised motion cues alone may be insufficient for reliable discrimination. For the UCF101 dataset, actions where human body joints are partially occluded - like videos of action classes "Rowing" or "Breast Stroke" - pose-only models show greater action classification accuracy. This could be caused by an introduction of significant noise in the flow field, which has less of an impact on the pose keypoints.

E. Performance and Implementation Considerations

PoseOFF provides a computationally efficient alternative to dense motion representations, retaining informative motion cues while avoiding the substantial overhead of full optical flow processing. Table II compares data size and extraction cost for pose-only, full optical flow, and PoseOFF representations. For a representative 1920×1080 video sequence, PoseOFF requires approximately 1060kB of data, compared to 60kB for pose-only and 1.66GB for full optical flow. This represents an orders-of-magnitude reduction in data relative to dense motion representations, while still incorporating meaningful motion information.

PoseOFF adds only a modest overhead to action inference time compared to pose-only inference, due to processing optical flow vectors embedded with keypoint features. As shown in Table III, it increases trainable parameters by an average of 1.88% and adds 5.3 ms and 9.35 ms to post-extraction inference time for InfoGCN++ and MS-G3D, respectively. These increases remain small relative to model complexity and are compatible with real-time operation.

Importantly, the modifications required to incorporate PoseOFF are limited to the embedding layer of each model, preserving the structure and efficiency of the underlying architecture. This enables straightforward integration into existing skeleton-based pipelines.

V. Discussion and Conclusion

The results demonstrate that incorporating localised motion cues through PoseOFF consistently improves early action recognition, particularly under partial observation. By conditioning motion extraction on pose, PoseOFF captures fine-grained dynamics—such as limb acceleration and object interaction—that are not represented in skeletal configurations alone, while avoiding the overhead of dense optical flow. This positions PoseOFF as an efficient middle ground between lightweight pose-based methods [3], [4] and computationally intensive motion-based approaches [5].

Per-class analysis shows that these gains are most pronounced for actions characterised by subtle, localised motion, including hand-object interactions and self-directed activities, where pose trajectories alone are insufficient for early discrimination. In contrast, actions dominated by large-scale global motion exhibit smaller or more variable improvements, suggesting that such dynamics are already well captured by skeletal representations.

From a systems perspective, PoseOFF provides a lightweight mechanism for enhancing existing skeleton-based models with targeted motion cues, requiring minimal architectural modification and modest computational overhead. This makes it well suited to low-latency and resource-constrained settings.

These properties are particularly relevant in human-robot interaction scenarios, where early and reliable interpretation of human intent directly impacts responsiveness, coordination, and safety [7], [1]. By enabling comparable performance with fewer observed frames, PoseOFF supports earlier prediction of human actions, facilitating more anticipatory

TABLE II: Data size and extraction speed comparison for pose-only, PoseOFF, and full optical flow representations.

Representation	Data Size	Extraction Time	FPS
Pose only	60 kB	713 ms	140
PoseOFF	1060 kB	720 ms	138
Full flow	1.66 GB	10.8 s	9

TABLE III: Model size and inference speed comparison for baseline and PoseOFF-augmented models.

Model	Variant	Parameters	Inference Time
InfoGCN++	Base	613,747	12.15 ms (82 fps)
	PoseOFF	620,448	17.44 ms (57 fps)
MS-G3D	Base	2,194,476	13.13 ms (76 fps)
	PoseOFF	2,204,420	22.48 ms (44 fps)
ST-GCN++	Base	516,454	4.43 ms (226 fps)
	PoseOFF	537,648	4.20 ms (238 fps)

Fig. 7: Qualitative examples illustrating how PoseOFF captures localised motion cues (e.g., rotation, divergence) around key joints, supporting improved recognition under occlusion, subtle motion, and scene clutter.

behaviours such as motion priming, adaptive planning, and proactive safety responses.

PoseOFF relies on the quality of pose estimation and local motion extraction, and may be less effective under severe occlusion or highly dynamic global motion. Future work will explore more efficient motion estimation strategies and real-time deployment to evaluate system-level performance in embodied interaction settings.

References

[1] P. A. Lasota, T. Fong, and J. A. Shah, "A Survey of Methods for Safe Human-Robot Interaction," *Found. Trends® Robot.*, vol. 5, no. 4, pp. 261–349, May 2017.

[2] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, Nov. 2018.

[3] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," Jan. 2018.

[4] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition With Directed Graph Neural Networks," in *2019 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 7904–7913.

[5] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," https://arxiv.org/abs/1406.2199v2, Jun. 2014.

[6] R. Girdhar and K. Grauman, "Anticipative Video Transformer," https://arxiv.org/abs/2106.02036v2, Jun. 2021.

[7] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACMIEEE Int. Conf. Hum.-Robot Interact. HRI*, Mar. 2013, pp. 301–308.

[8] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *2011 Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1036–1043.

[9] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging LSTMs to Anticipate Actions Very Early," https://arxiv.org/abs/1703.07023v3, Mar. 2017.

[10] Y. Kong and Y. Fu, "Human Action Recognition and Prediction: A Survey," *Int J Comput Vis*, vol. 130, no. 5, pp. 1366–1401, May 2022.

[11] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting Skeleton-based Action Recognition," in *2022 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 2959–2968.

[12] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training," Oct. 2022.

[13] J. Cai, N. Jiang, X. Han, K. Jia, and J. Lu, "JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition," in *2021 IEEE Winter Conf. Appl. Comput. Vis. WACV*. Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 2734–2743.

[14] Z. Huang, X. Liu, and Y. Kong, "H-MoRe: Learning Human-centric Motion Representation for Action Analysis," in *2025 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*. Nashville, TN, USA: IEEE, Jun. 2025, pp. 22 702–22 713.

[15] S. Chi, H.-G. Chi, Q. Huang, and K. Ramani, "InfoGCN++: Learning Representation by Predicting the Future for Online Skeleton-Based Action Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 1, pp. 514–528, Jan. 2025.

[16] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," May 2020.

[17] H. Duan, J. Wang, K. Chen, and D. Lin, "PYSKL: Towards Good Practices for Skeleton Action Recognition," May 2022.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Adv. Neural Inf. Process. Syst.*, 2019.

[19] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," Aug. 2020.

[20] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss," in *2022 IEEECVF Conf. Comput. Vis. Pattern Recognit. Workshop CVPRW*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 2636–2645.

[21] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 1010–1019.

[22] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[23] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *ArXiv Prepr. ArXiv12120402*, 2012.