Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sport Classification from Multi-Player Trajectories with Set-over-Time Aggregation

Sawet Somnugpong

sawet_s@kpru.ac.th

Kamphaeng Phet Rajabhat University

Research Article

Keywords: sport classification, multi-player trajectories, set-over-time aggregation, permutation-invariant modeling

Posted Date: April 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9214484/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sport Classification from Multi-Player Trajectories with Set-over-Time Aggregation

Sawet Somnugpong[1*]

[1*]Computer Technology Program, Faculty of Industrial Technology, Kamphaeng Phet Rajabhat University, Kamphaeng Phet 62000, Thailand.

Corresponding author(s). E-mail(s): sawet_s@kpru.ac.th;

Abstract

Sport classification is commonly studied from RGB video, but multi-player trajectories also provide a compact representation of movement, spacing, and interaction patterns. This paper investigates sport classification in a trajectory-only setting and introduces a set-over-time formulation for multi-player inputs. Each clip is represented as a sequence of frames, where each frame contains an unordered set of visible player trajectories, and the method is evaluated on a four-sport classification task built from MultiSports actor tubes. The proposed model first encodes the player set in each frame with permutation-invariant aggregation and then performs temporal aggregation for clip-level prediction. This design preserves frame-level multi-player structure before temporal summarization. Experiments compare the proposed approach with three compact trajectory baselines: mean pooling, mean-plus-standard-deviation pooling, and GRU-based aggregation. The proposed model achieves the best validation macro-F1 of 0.8793 and test macro-F1 of 0.8614, outperforming the strongest baseline by 0.0386 test macro-F1. Ablation results further show that entity weighting, temporal variability modeling, time attention, and the joint use of position and motion cues all contribute to performance. Error analysis indicates that trajectory-only recognition is effective, but confusion remains among team sports with partially similar collective motion patterns. Overall, the results show that set-over-time aggregation is an effective approach for sport classification from multi-player trajectories.

Keywords: sport classification, multi-player trajectories, set-over-time aggregation, permutation-invariant modeling

1 Introduction

Sport classification is commonly studied from RGB video, where models rely on appearance, scene context, and temporal dynamics. However, in many sports, an important part of clip semantics is also encoded in how multiple players are spatially arranged and how their movement evolves collectively over time. This motivates the use of multi-player trajectories as a compact representation that emphasizes motion, spacing, and interaction while reducing reliance on texture and background cues.

A key challenge is that the visible players in each frame do not form an ordered sequence. Instead, they constitute an unordered set whose size may vary because of occlusion, truncation, and tracking quality. Standard temporal models such as GRU, LSTM, and Bi-LSTM are effective for sequence modeling [1–3], but they do not resolve the permutation ambiguity within each frame. For this reason, permutation-invariant

aggregation is a more suitable inductive bias for frame-level player observations, as also suggested by prior work on set encoding and sports context modeling [4, 5].

Trajectory-based representations have shown value in sports analytics and related multi-agent problems, including ball inference, forecasting, and interaction modeling [5–10]. Nevertheless, direct sport classification from multi-player trajectories remains relatively underexplored, particularly under a formulation that explicitly respects the unordered structure of within-frame player observations.

Motivated by this gap, this paper studies sport classification from multi-player trajectories using a set-over-time formulation. The proposed model first encodes the unordered player set in each frame through permutation-invariant aggregation and then summarizes the resulting frame sequence temporally. A four-sport trajectory classification task is constructed from MultiSports actor tubes [11] to evaluate this design. The model is compared against compact trajectory baselines based on mean pooling, mean-plus-standard-deviation pooling, and GRU-based aggregation. Results show that preserving frame-level multi-player structure before temporal summarization leads to stronger sport discrimination from trajectories alone. Ablation analyses further confirm the importance of entity weighting, temporal variability modeling, time attention, and the joint use of position and motion cues.

2 Related Work

2.1 Trajectory modeling in sports and multi-agent scenes

Trajectory learning has been widely studied in multi-agent environments because trajectories encode interaction, intent, and scene dynamics in a compact form. In sports analytics, Kim et al. showed that multi-agent player context can improve ball trajectory inference through set-based context encoding and hierarchical temporal modeling [5]. TranSPORTmer further demonstrated that unified transformer-style modeling can support forecasting, imputation, and state understanding for multi-agent sports trajectories [6]. These studies show that trajectory-first modeling is useful for sports reasoning, but they do not directly address clip-level sport classification.

Beyond sports, several influential methods addressed human and agent trajectories in social settings. Social LSTM introduced recurrent interaction modeling through pooled neighboring trajectories, Social GAN combined trajectory prediction with adversarial training for socially acceptable forecasts, and Trajectron/Trajectron++ extended trajectory reasoning with graph-structured and probabilistic formulations [7–10]. Although these approaches were not designed for sport recognition, they highlight the importance of multi-agent context and variable-size interaction structure. The setting studied here differs in that it focuses on clip-level sport classification and emphasizes a compact representation that preserves within-frame structure before temporal aggregation.

2.2 Set-based and sequence-based modeling for unordered player observations

A key difficulty in this problem is that the visible players within a frame form an unordered set. Any arbitrary player ordering can change across clips or even between adjacent frames, making order-sensitive encoders brittle. Permutation-invariant processing and symmetric aggregation over unordered observations have proved effective in point-set learning and sports context modeling, as illustrated by PointNet and set-based ball-trajectory inference [4, 5]. These ideas offer a principled alternative to imposing a fixed player order.

Temporal modeling remains essential because sport identity is expressed over time. Recurrent units such as LSTM, GRU, and Bi-LSTM are standard tools for sequence learning [1–3]. In video understanding, stronger temporal or global aggregation has been explored through segmental modeling, non-local operations, action transformers, hierarchical transformers, and self-supervised video encoders [12–16]. The proposed formulation connects these two lines of work: each frame is first treated as a set of players, and only then is temporal aggregation applied across frame-level embeddings. Compared with sequence-first baselines, this design keeps the representation aligned

with the unordered nature of within-frame player observations.

2.3 Sports video benchmarks and relation to the present setting

A large body of work on action and sports understanding has been driven by RGB video benchmarks and increasingly powerful video backbones. Datasets such as FineGym, SoccerNet, SoccerNet-v2, and MultiSports have supported progress in fine-grained sports analysis and sports event understanding [11, 17–19]. Corresponding model families include ResNet-style CNNs, C3D, Temporal Segment Networks, I3D, SlowFast, Non-local Networks, Swin Transformer, and VideoMAE [12, 13, 15, 16, 20–23]. Sports video pipelines that produce player tubes also rely on detection and tracking components, where Faster R-CNN, YOLO, DETR, SORT, and ByteTrack are representative building blocks [24–28]. These methods substantially improved appearance-rich video recognition and visual localization.

This work is related to this literature only in the sense that it uses sports video annotations as a source of trajectory data. The focus here is not to compete with RGB video models, but to study what can be recognized from player trajectories alone. MultiSports is therefore used as an evaluation source for constructing a trajectory-only four-sport classification task, while the main contribution remains the set-over-time modeling formulation rather than the benchmark itself.

3 Methodology

Figure 1 summarizes the proposed pipeline, from trajectory clip construction to set-over-time encoding and final sport prediction.

3.1 Problem formulation

Given a trajectory clip, the goal is to predict its sport label from multi-player motion alone. Each clip is represented as a variable-length sequence of frames, and each frame contains a variable number of visible players. Let

$$\mathcal{X} = \{\mathcal{S}_t\}_{t=1}^{T}, \qquad \mathcal{S}_t = \{\mathbf{x}_{t,p}\}_{p=1}^{P_t}, \tag{1}$$

where $\mathcal{S}_t$ is the set of player observations at frame t, P_t is the number of valid entities in that frame,

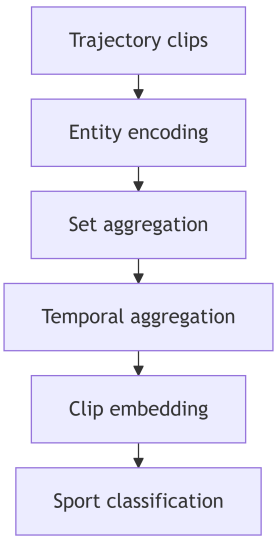


Fig. 1 Overview of the proposed set-over-time trajectory classification pipeline.

and $\mathbf{x}_{t,p} \in \mathbb{R}^D$ is the feature vector of entity p. In this implementation, $D = 12$. The classification target is a four-way sport label

$$y \in \{\text{aerobic_gymnastics}, \text{volleyball}, \text{football}, \text{basketball}\}. \tag{2}$$

The key modeling issue is that players within a frame do not form a naturally ordered sequence. Each frame is therefore modeled as an unordered set and each clip as a sequence of sets. This motivates a set-over-time architecture: permutation-invariant aggregation is applied over players within each frame, and temporal aggregation is applied only afterward across frame-level embeddings, following the general rationale of set-based modeling for unordered observations [4, 5].

3.2 Trajectory sample construction

The trajectory classification task is derived from the dense actor tubes provided by MultiSports [11]. Each training sample is built from one anchor tube and a small set of temporally overlapping context tubes from the same video. The anchor tube defines the time axis of the sample, and all context entities are aligned to the anchor frames. This gives a clip-level representation that preserves the motion of the primary actor while retaining nearby multi-player context.

For each anchor tube, up to $P_{\max} - 1$ additional context tubes with the largest temporal overlap with the anchor are selected. In the experiments, $P_{\max} = 8$, so each sample contains at most one anchor entity and seven context entities. Samples shorter than 12 frames are discarded. This procedure converts the original actor-tube annotations into a trajectory-only sport classification task in which each clip is a tensor of shape $T \times P \times D$ together with validity masks over time and entities.

3.3 Per-entity trajectory features

For each valid player box at frame t, a 12-dimensional feature vector is extracted

$$\mathbf{x}_{t,p} = [x_1, y_1, x_2, y_2, c_x, c_y, w, h, v_x, v_y, s, a]. \quad (3)$$

The first eight components describe box geometry: top-left and bottom-right coordinates, center location (c_x, c_y), width w, and height h. The motion terms are computed along the anchor-aligned timeline as

$$\begin{aligned} v_x &= c_x^{(t)} - c_x^{(t-1)}, \\ v_y &= c_y^{(t)} - c_y^{(t-1)}, \\ s &= \sqrt{v_x^2 + v_y^2}. \end{aligned} \quad (4)$$

and the area term is defined as

$$a = w \cdot h. \quad (5)$$

If an entity is not visible at a particular anchor frame, its feature vector is zero padded and excluded from aggregation through a binary validity mask $m_{t,p} \in \{0, 1\}$. This representation lets the model use both spatial configuration and temporal dynamics while remaining compact and fully trajectory-based. The ablation study also evaluates two restricted feature groups: position-only features (c_x, c_y, w, h) and motion-only features (v_x, v_y, s).

3.4 Set-over-time trajectory encoder

The proposed model first encodes the unordered player set within each frame and then aggregates the resulting frame embeddings over time. Given an entity feature vector $\mathbf{x}_{t,p}$, an MLP encoder $\phi(\cdot)$ produces a hidden representation

$$\mathbf{h}_{t,p} = \phi(\mathbf{x}_{t,p}). \quad (6)$$

Frame-level set encoding.

To allow different entities to contribute unequally, an entity importance score is computed

$$e_{t,p} = s_{\text{ent}}(\mathbf{h}_{t,p}), \quad (7)$$

mask invalid entities, and normalize the remaining scores with a softmax over players:

$$\alpha_{t,p} = \frac{m_{t,p} \exp(e_{t,p})}{\sum_j m_{t,j} \exp(e_{t,j})}. \quad (8)$$

Using these weights, a permutation-invariant weighted set mean is formed

$$\boldsymbol{\mu}_t^E = \sum_p \alpha_{t,p} \mathbf{h}_{t,p}, \quad (9)$$

and a weighted set standard deviation

$$\boldsymbol{\sigma}_t^E = \sqrt{\sum_p \alpha_{t,p} \left(\mathbf{h}_{t,p} - \boldsymbol{\mu}_t^E\right)^2}. \quad (10)$$

The final frame embedding is obtained by concatenating these two statistics and projecting them with another MLP $\rho(\cdot)$:

$$\mathbf{g}_t = \rho\left([\boldsymbol{\mu}_t^E, \boldsymbol{\sigma}_t^E]\right). \quad (11)$$

Temporal aggregation.

The sequence of frame embeddings $\{\mathbf{g}_t\}_{t=1}^T$ is then aggregated over time. A temporal score is computed for each valid frame,

$$u_t = s_{\text{time}}(\mathbf{g}_t), \quad (12)$$

and apply a masked softmax to obtain temporal weights

$$\beta_t = \frac{\tau_t \exp(u_t)}{\sum_k \tau_k \exp(u_k)}, \quad (13)$$

where $\tau_t \in \{0, 1\}$ indicates whether frame t is valid in the padded batch. The clip-level temporal mean and standard deviation are

$$\boldsymbol{\mu}^T = \sum_t \beta_t \mathbf{g}_t, \quad (14)$$

$$\boldsymbol{\sigma}^T = \sqrt{\sum_t \beta_t \left(\mathbf{g}_t - \boldsymbol{\mu}^T\right)^2}. \quad (15)$$

Algorithm 1 Set-over-Time trajectory classification

Require: trajectory clip $\mathcal{X} = \{\mathcal{S}_t\}_{t=1}^{T}$, entity masks $\{m_{t,p}\}$, frame masks $\{\tau_t\}$

for $t = 1$ to T **do**
  for each valid entity p in frame t **do**
    Encode entity feature: $\mathbf{h}_{t,p} \leftarrow \phi(\mathbf{x}_{t,p})$
    Compute entity score: $e_{t,p} \leftarrow s_{\text{ent}}(\mathbf{h}_{t,p})$
  end for
  Normalize masked entity scores to obtain $\alpha_{t,p}$
  Compute weighted set statistics $\boldsymbol{\mu}_t^E$ and $\boldsymbol{\sigma}_t^E$
  Form frame embedding $\mathbf{g}_t \leftarrow \rho([\boldsymbol{\mu}_t^E, \boldsymbol{\sigma}_t^E])$
end for
for $t = 1$ to T **do**
  Compute temporal score $u_t \leftarrow s_{\text{time}}(\mathbf{g}_t)$
end for
Normalize masked temporal scores to obtain β_t
Compute temporal statistics $\boldsymbol{\mu}^T$ and $\boldsymbol{\sigma}^T$
Form clip embedding $\mathbf{z} \leftarrow [\boldsymbol{\mu}^T, \boldsymbol{\sigma}^T]$
Predict sport label $\hat{y} \leftarrow \text{Classifier}(\mathbf{z})$

The final clip embedding is

$$\mathbf{z} = [\boldsymbol{\mu}^T, \boldsymbol{\sigma}^T], \tag{16}$$

which is fed to a classification head to predict the sport label.

This design has two advantages for this task. First, the frame encoder is permutation invariant with respect to player order, which matches the structure of multi-player observations and is consistent with prior set-based encoders for unordered inputs [4, 5]. Second, the model preserves frame-level context before temporal summarization, unlike sequence-first baselines that average player information too early.

3.5 Training objective and analysis variants

The model is trained with standard cross-entropy loss over the four sport classes:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{4} y_c \log \hat{y}_c. \tag{17}$$

Variable-length clips and different numbers of entities are handled by zero padding together with temporal and entity masks, ensuring that padded positions do not affect any set or temporal aggregation step.

For comparison, three trajectory baselines are evaluated: mean pooling, mean-plus-standard-deviation pooling, and a GRU-based temporal model [2]. All three baselines first collapse the entity dimension by simple averaging and then differ only in the way they summarize the resulting frame sequence. Five ablation variants of the proposed model are also defined: without entity weighting, without temporal standard deviation, without time attention, position-only input, and motion-only input. These variants are used to test whether the gains come from the set-over-time formulation itself or from specific design choices inside the model.

4 Experiments and Results

4.1 Experimental Setup

The proposed set-over-time formulation is evaluated on the four-sport trajectory classification task described in Section 3.2. The task is built from MultiSports actor tubes [11] and contains four sport categories: aerobic_gymnastics, volleyball, football, and basketball. The resulting split contains 11,671 training clips, 2,030 validation clips, and 4,786 test clips. The test set is moderately imbalanced, with 1,400 aerobic_gymnastics clips, 1,155 volleyball clips, 1,289 football clips, and 942 basketball clips.

The proposed model is compared against three compact trajectory baselines: mean pooling, mean-plus-standard-deviation pooling, and a GRU-based temporal model [2]. All methods use the same trajectory samples and the same per-entity features described in Section 3.3. In the implementation, clips are trained for 18 epochs with batch size 32, hidden dimension 128, dropout 0.2, learning rate 10^{-3}, and weight decay 10^{-4}. Both classification accuracy and macro-F1 are reported. Because the class frequencies are not perfectly balanced, macro-F1 is treated as the primary reporting metric. Best checkpoints are selected by validation accuracy, with validation macro-F1 used as a tie-breaker.

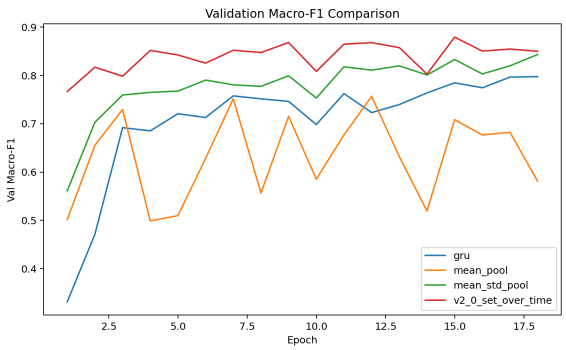


Fig. 2 Validation macro-F1 comparison for the main models.

4.2 Main Comparison Results

Table 1 summarizes the main comparison. The proposed set-over-time model achieves the strongest results on both validation and test sets, reaching a validation macro-F1 of 0.8793 and a test macro-F1 of 0.8614. The best baseline is mean-plus-standard-deviation pooling, which achieves a test macro-F1 of 0.8227. The absolute gain of the proposed model over this strongest baseline is therefore 0.0386 test macro-F1, indicating that preserving frame-level multi-player structure before temporal aggregation is beneficial for sport classification from trajectories alone.

Figure 2 provides a validation macro-F1 comparison across the four methods. The ranking is consistent with the table, with the proposed model clearly outperforming the simpler pooling baselines and the GRU baseline. This pattern suggests that a sequence-only formulation is not sufficient by itself; the representation also benefits from modeling the within-frame player collection as an unordered set, which is consistent with prior work on permutation-invariant set encoding [4, 5].

The learning curves in Figure 3 show that the proposed model converges stably. Training loss decreases steadily, while validation macro-F1 rises rapidly during the early epochs and reaches its best value around epoch 15. After that point, the validation curves become slightly more oscillatory while the training objective continues to improve, suggesting mild overfitting in later epochs. Even so, the overall optimization behavior remains stable and leads to the best generalization among all compared methods.

4.3 Ablation Study

The contribution of the main architectural components is further analyzed through ablation experiments. Table 2 reports the core architecture ablations, where the full model is compared with variants that remove entity weighting, temporal standard deviation, or time attention. In this separate ablation run, the full model achieves the best test macro-F1 of 0.8679. Removing temporal standard deviation reduces test macro-F1 to 0.8573, and removing entity weighting reduces it to 0.8608. These drops indicate that both reliability-aware set aggregation and temporal variability cues contribute to the final representation. The variant without time attention obtains 0.8628 test macro-F1, showing that emphasizing informative time steps also provides a measurable benefit.

The feature-group ablation in Table 3 further shows that neither positional features alone nor motion features alone are sufficient. Both the position-only and motion-only variants collapse to a test macro-F1 of 0.1061, far below the full model. Their identical results further suggest a degenerate single-class solution, indicating that neither cue alone provides sufficient discriminative information for this task. This result highlights that the task depends on the interaction between spatial configuration and temporal dynamics rather than on either source of information in isolation.

Overall, the ablation study supports the design choice of combining entity weighting, temporal dispersion modeling, time attention, and joint position–motion features in a unified set-over-time architecture.

4.4 Error Analysis

Figure 4 shows the test confusion matrix of the proposed model. The easiest class is aerobic_gymnastics, for which the model correctly recognizes 1320 out of 1400 test clips. Football is also recognized reliably, with 1215 correct predictions out of 1289. By contrast, volleyball is the most difficult class, with 876 correct predictions out of 1155. Basketball lies in between, with 754 correct predictions out of 942. The largest confusions occur between volleyball and football, and between basketball and volleyball. In particular, 172 volleyball clips are misclassified as football, 157 basketball clips are misclassified as volleyball,

Table 1 Main comparison on the four-sport trajectory classification task.

Model	Best Epoch	Val Macro-F1	Test Macro-F1	Test Acc
Mean pooling	7	0.7517	0.6609	0.6732
GRU	18	0.7975	0.7422	0.7443
Mean+Std pooling	18	0.8432	0.8227	0.8282
Proposed	**15**	**0.8793**	**0.8614**	**0.8702**

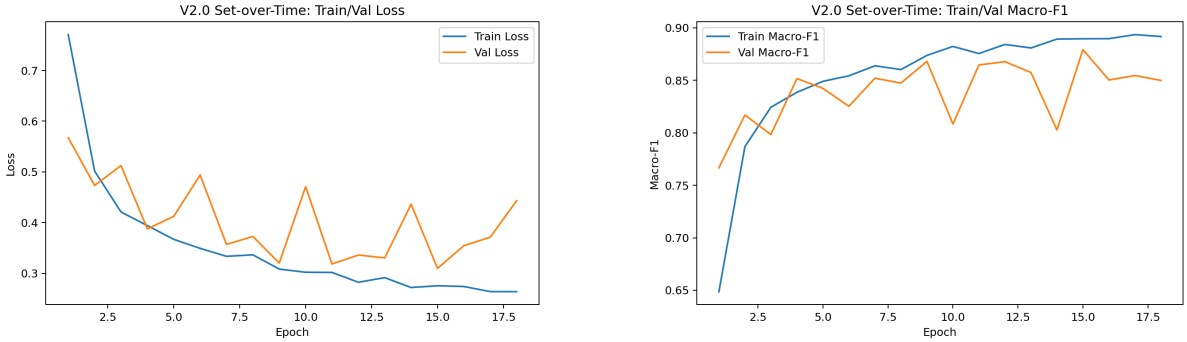


Fig. 3 Training dynamics of the proposed model. Left: loss curves. Right: macro-F1 curves.

Table 2 Core ablation results for the proposed architecture.

Variant	Best Epoch	Test Macro-F1	Test Acc
w/o entity weighting	12	0.8608	0.8711
w/o temporal std	18	0.8573	0.8644
w/o time attention	11	0.8628	0.8709
Full model	**17**	**0.8679**	**0.8761**

Table 3 Feature-group ablation results.

Variant	Best Epoch	Test Macro-F1	Test Acc
Position only	1	0.1061	0.2693
Motion only	1	0.1061	0.2693
Full model	**17**	**0.8679**	**0.8761**

and 102 volleyball clips are misclassified as basketball. This pattern suggests that team sports with partially similar collective motion patterns remain harder to separate than highly distinctive motion classes such as aerobic gymnastics. It also indicates that further gains may come from stronger modeling of inter-player coordination and finer temporal structure in ambiguous team-sport clips, as explored more broadly in recent sports and multi-agent trajectory models [5, 6].

5 Discussion

The results indicate that the main advantage of the proposed approach comes from matching the model structure to the data structure. In this task, the informative unit at each time step is not an ordered list of players but a multi-player configuration. Treating frame observations as sets before temporal aggregation therefore preserves interaction context that would otherwise be weakened by early pooling or sequence-first modeling. The gain over compact baselines suggests that sport identity in trajectory-only settings is carried not only by average movement, but by how player configurations are organized and evolve collectively over time.

The ablation study further clarifies that the improvement is not attributable to a single design

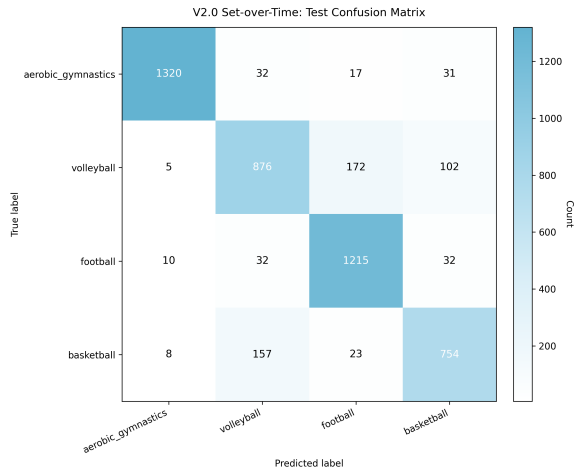


Fig. 4 Test confusion matrix of the proposed model.

choice. Entity weighting helps the model suppress less informative or noisier player observations, which is especially useful when visibility and tracking quality vary across clips. Temporal variability modeling contributes because sports are distinguished not only by typical positions or motion magnitudes, but also by how stable or dynamic these patterns remain across a clip. Time attention provides an additional benefit by emphasizing moments in which class-specific motion structure is most visible. Taken together, these findings suggest that trajectory-based sport classification benefits from jointly modeling salience, variability, and temporal importance rather than relying on a single pooled descriptor.

The remaining errors are also informative. Most confusions occur among volleyball, football, and basketball, which implies that trajectory-only representations capture broad collective motion structure well, but are less sufficient when different team sports exhibit partially similar short-term spatial layouts or transition patterns. In contrast, aerobic gymnastics is easier to separate because its movement organization is more distinctive. This pattern suggests that the current model is effective for compact motion-based recognition, but that further gains for closely related team sports will likely require richer interaction modeling or multimodal cues such as court context, ball information, or appearance features.

6 Conclusion

This paper studied sport classification from multi-player trajectories in a trajectory-only setting. Using a four-sport benchmark derived from MultiSports actor tubes, this study introduced a compact set-over-time architecture that treats the visible players in each frame as an unordered set before temporal aggregation.

The proposed model achieved the best overall performance among the compact trajectory baselines and the ablation results confirmed the benefit of entity weighting, temporal variability modeling, time attention, and the joint use of position and motion cues. These findings support set-first, time-second aggregation as an effective design for multi-player trajectory classification.

The error analysis also indicates that trajectory-only recognition remains challenging for team sports with similar collective movement patterns. Future work will explore richer interaction modeling and multimodal extensions that combine trajectory and appearance information.

Declarations

Funding

Not applicable.

Data availability

The MultiSports dataset used in this study is publicly available from its original source. The derived trajectory clips, preprocessed features, and experimental splits generated for this study are available from the corresponding author on reasonable request.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997) https://doi.org/10.1162/neco.1997.9.8.1735

[2] Cho, K., Merri"enboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1724–1734 (2014). https://doi.org/10.3115/v1/d14-1179

[3] Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997) https://doi.org/10.1109/78.650093

[4] Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017). https://doi.org/10.1109/cvpr.2017.16

[5] Kim, H., Choi, H.-J., Kim, C.J., Yoon, J., Ko, S.-K.: Ball trajectory inference from multi-agent sports contexts using set transformer and hierarchical Bi-LSTM. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2023). https://doi.org/10.1145/3580305.3599779

[6] Capellera, G., Ferraz, L., Rubio, A., Agudo, A., Moreno-Noguer, F.: TranSPORTmer: A holistic approach to trajectory understanding in multi-agent sports. In: Proceedings of the Asian Conference on Computer Vision, pp. 333–352 (2024). https://doi.org/10.1007/978-981-96-0901-7_20

[7] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–971 (2016). https://doi.org/10.1109/cvpr.2016.110

[8] Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2255–2264 (2018). https://doi.org/10.1109/cvpr.2018.00240

[9] Ivanovic, B., Harrison, J., Sharma, A., Chen, M., Li, C., Chou, S., Behbahani, F., Yue, Y., Pavone, M.: The Trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019). https://doi.org/10.1109/iccv.2019.00246

[10] Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: Proceedings of the European Conference on Computer Vision, pp. 683–700 (2020). https://doi.org/10.1007/978-3-030-58523-5_40

[11] Li, Y., Chen, L., He, R., Wang, Z., Wu, G., Wang, L.: MultiSports: A multi-person video dataset of spatio-temporally localized sports actions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13536–13545 (2021). https://doi.org/10.1109/iccv48922.2021.01328

[12] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of the European Conference on Computer Vision, pp. 20–36 (2016). https://doi.org/10.1007/978-3-319-46484-8_2

[13] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018). https://doi.org/10.1109/cvpr.2018.00813

[14] Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253 (2019). https://doi.org/10.1109/cvpr.2019.00033

[15] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In:

Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021). https://doi.org/10.1109/iccv48922.2021.00986

[16] Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems (2022). https://doi.org/10.52202/068431-0732

[17] Shao, D., Zhao, Y., Dai, B., Lin, D.: FineGym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2616–2625 (2020). https://doi.org/10.1109/cvpr42600.2020.00269

[18] Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: SoccerNet: A scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2018). https://doi.org/10.1109/cvprw.2018.00223

[19] Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasr-Esfahani, E., Ghanem, B., Van Droogenbroeck, M.: SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 2215–2224 (2021). https://doi.org/10.1109/cvprw53098.2021.00508

[20] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.1109/cvpr.2016.90

[21] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015). https://doi.org/10.1109/iccv.2015.510

[22] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the Kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017). https://doi.org/10.1109/cvpr.2017.502

[23] Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019). https://doi.org/10.1109/iccv.2019.00630

[24] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015). https://doi.org/10.1109/tpami.2016.2577031

[25] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016). https://doi.org/10.1109/cvpr.2016.91

[26] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision, pp. 213–229 (2020). https://doi.org/10.1007/978-3-030-58452-8_13

[27] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: Proceedings of the IEEE International Conference on Image Processing, pp. 3464–3468 (2016). https://doi.org/10.1109/icip.2016.7533003

[28] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-object tracking by associating every detection box. In: Proceedings of the European Conference on Computer Vision, pp. 1–21 (2022). https://doi.org/10.1007/978-3-031-20047-2_1