IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1173-1184
Journal homepage: http://iieta.org/journals/ts

Image Recognition-Based Automatic Detection of Rule Violations in Soccer Match Videos

Qi Xie[1*], Tianyu Sun[2]

[1] Department of Physical Education, Baoji University of Arts and Sciences,Baoji, Shaanxi 721013, China
[2] General Quality Characteristics Research Center, Xi'an Institute of Modern Control Technology, Xi'an 710065, China

Corresponding Author Email: moray1981@163.com

https://doi.org/10.18280/ts.430308

Received: 9 November 2025
Revised: 28 May 2026
Accepted: 6 June 2026
Available online: 30 June 2026

Keywords:
video action recognition, graph convolutional network, multi-view feature fusion, interpretable visual recognition, intelligent sports image analysis

ABSTRACT

Fine-grained human interaction detection represents a fundamental research direction in video action understanding. In competitive sports scenarios, high-speed motion, dense multi-player occlusion, and subtle inter-class behavioral differences impose stringent requirements on detection accuracy and result interpretability. Existing purely data-driven behavior detection approaches encounter substantial limitations in effectively incorporating domain-specific knowledge of competition rules, resulting in inadequate generalization to rare foul events, inflexible multi-view fusion, and insufficient quantitative evidence to support decision-making. Consequently, the requirements of professional referee-assisted decision systems cannot be satisfactorily fulfilled. To address these challenges, a three-module serial-feedback coupled multi-view framework for automatic soccer rule violation detection was proposed, enabling end-to-end output from raw video sequences to foul categories, severity assessment, and interpretable decision evidence. Competition rules were first formalized as computable semantic triggers, employing rule-constrained spatiotemporal graph convolution to achieve explicit domain-knowledge guidance over visual feature aggregation. A dynamically weighted dual-branch spatiotemporal query decoding architecture was then constructed to adaptively allocate multi-view fusion weights, emulating the multi-view verification strategy adopted during referee decision-making. Furthermore, a geometric potential field model was established to quantitatively assess foul severity through the evolution of multi-component energy curves. Closed-loop collaboration among the modules was achieved through three feedback pathways, iteratively optimizing the detection performance. Experimental results demonstrated that the proposed framework consistently outperformed representative baseline methods in detection accuracy, officiating consistency, and few-shot generalization capability, while simultaneously providing physically interpretable visual evidence to support decision-making. This study enriches the methodological pathways for deeply integrating symbolic knowledge with visual features and offers technical support for intelligent image analysis in competitive sports.

1. INTRODUCTION

Fine-grained human interaction detection constitutes a fundamental research direction in video action understanding [1-3] and has been widely recognized as having substantial application value in intelligent sports analytics, video-assisted officiating, motion capture, and athletic training analysis. As a representative multi-player competitive sport characterized by high-speed physical confrontation, soccer presents distinctive challenges for visual detection tasks [4-6]. These challenges arise from the dense spatial distribution of players accompanied by dynamic occlusions, the rapid execution and short duration of actions, the subtle visual differences between legal and rule-violating actions, and the strong dependence of officiating decisions on domain-specific knowledge and contextual semantics. Consequently, this scenario has been regarded as a representative benchmark for evaluating the performance of fine-grained action detection algorithms. Current video assistant referee systems continue to rely primarily on the manual review of multi-view video recordings as the principal decision-making process [7-9]. As a result, limitations associated with low review efficiency and the susceptibility of officiating decisions to subjective judgment remain unavoidable, highlighting the urgent need for automated image recognition techniques to provide technical support. Meanwhile, existing general-purpose video action detection models have largely been developed for generic action categories [10-12], making it difficult for them to simultaneously satisfy the stringent requirements for both detection accuracy and result interpretability demanded by professional sports officiating. Consequently, these models cannot be directly adapted to competitive sports scenarios governed by strict rule constraints. To address this challenge, the task of automatic soccer rule violation detection is investigated. The resulting methodology is expected to enrich the technical paradigm for the deep integration of symbolic domain knowledge with visual representations while providing a practical image recognition solution for intelligent

officiating systems in competitive sports.

From the perspective of existing technical frameworks, four common technical bottlenecks continue to constrain fine-grained rule violation detection in competitive sports scenarios. First, the integration of domain knowledge with visual representations remains insufficient [13, 14]. Existing action detection models have predominantly been developed under purely data-driven training paradigms, making it difficult for domain-specific prior knowledge, such as competition rules, to be explicitly incorporated into the feature aggregation process. As a consequence, limited generalization to rare foul categories has been achieved, semantically traceable evidence supporting model decisions has been lacking, and the interpretability of detection results has remained inadequate for the reliability requirements of professional officiating. Second, existing multi-view feature fusion mechanisms lack dynamic adaptability. Most current multi-view action recognition approaches [15, 16] have relied on static fusion strategies, including direct feature concatenation and fixed-weight feature aggregation. Consequently, the contribution of individual views cannot be adaptively adjusted according to the spatial location of an action or the degree of player occlusion. As a result, detection accuracy deteriorates substantially when local viewpoints are compromised, making it difficult for existing methods to emulate the multi-view verification strategy employed by human referees. Third, quantitative support for fine-grained decision interpretation remains insufficient [17, 18]. Most existing approaches have been limited to the output of discrete foul category labels, whereas a physically grounded quantitative representation of action hazardousness and rule violation severity has not been established. Consequently, hierarchical severity assessment cannot be achieved, and visual, traceable evidence to support final officiating decisions cannot be effectively provided. Fourth, collaborative closed-loop interaction among system modules remains absent [19, 20]. In existing detection frameworks, pose estimation, action classification, and result generation have generally been organized as isolated sequential stages. Consequently, high-level semantic information cannot be propagated back to low-level visual representations, making iterative performance optimization through bidirectional interaction difficult to achieve.

To address the aforementioned technical bottlenecks, a three-module serial-feedback coupled multi-view framework for automatic soccer rule violation detection is proposed. The core technical innovations are reflected in four aspects. First, a rule-guided spatiotemporal graph convolutional network is proposed, in which soccer competition rules are formalized as computable semantic triggers. Through the explicit modulation of graph attention weights by rule terms, domain knowledge directly provides guidance for visual feature aggregation, thereby enhancing both feature discriminability and generalization capability under few-shot scenarios. Second, a dynamically weighted dual-branch spatiotemporal query decoding framework is constructed. Two categories of learnable query vectors, corresponding to action category and severity assessment, are designed, and a dynamically view-weighted cross-attention mechanism is introduced to adaptively allocate fusion weights across different viewpoints, thereby improving detection robustness under occlusion. Third, a geometric potential field model based on human body keypoints is established. Three interpretable potential energy components, namely motion potential, contact potential, and hazardous posture potential, are defined, and quantitative severity assessment is achieved through the evolution of energy curves across consecutive frames, thereby providing physically meaningful visual explanations for officiating decisions. Fourth, a three-module serial-feedback coupling mechanism is established. Bidirectional collaboration among the three modules is achieved through rule-feature bias initialization, classification-prior dynamic weighting, and energy-conflict backward calibration, resulting in an end-to-end closed-loop optimization framework for rule violation detection.

The remainder of this study is organized as follows. Section 2 presents the proposed three-module coupled detection framework in detail, with the technical implementation of each module and the collaborative mechanisms among the modules described sequentially. Section 3 introduces the experimental datasets and evaluation metrics. Comprehensive performance is validated through extensive comparative and ablation experiments, followed by dedicated analyses of generalization capability, interpretability, and engineering feasibility. Sections 4 and 5 summarize the major findings, discuss the current limitations of the proposed framework, and outline directions for future research.

2. METHODOLOGY

2.1 Overview of the proposed framework

A three-module serial-feedback coupled multi-view framework for automatic soccer rule violation detection is proposed. Synchronized multi-view match video sequences are taken as the input, and three types of outputs are generated in an end-to-end manner, including foul category classification, severity assessment, and interpretable energy evolution curves. The overall architecture of the proposed framework is illustrated in Figure 1. During forward inference, the framework sequentially passes through three core modules: a rule-guided spatiotemporal graph convolutional network, a dynamically weighted dual-branch spatiotemporal query fusion network, and a geometric potential field-based fine-grained assessment module. In the first module, competition rules are formalized as computable semantic constraints and embedded into the feature aggregation process of the spatiotemporal graph convolutional network, thereby generating frame-level semantic representations enriched with rule-based prior knowledge. In the second module, adaptive fusion of multi-view features is performed through dual-branch learnable query embeddings, and preliminary predictions of foul category and severity assessment are produced. In the third module, a multi-component geometric potential field is constructed based on high-precision human body keypoints, through which quantitative severity assessment and interpretable visual explanations are generated.

Unlike conventional sequential detection architectures, a closed-loop optimization mechanism is established through three feedback pathways connecting the three modules. Rule-constrained features are utilized to provide initialization biases for the query vectors, classification predictions are employed to dynamically adjust the weights of the potential energy field, and backward calibration is triggered whenever inconsistencies arise between the energy-based quantitative assessment and the classification results. Through this design,

bidirectional interaction among low-level visual representations, mid-level semantic representations, and high-level rule-based knowledge is achieved. Consequently, detection accuracy is improved while the consistency and interpretability of the final decision are simultaneously ensured, thereby satisfying the stringent requirements for both accuracy and reliability in professional sports officiating.

2.2 Rule-guided spatiotemporal graph convolutional network

The objective of this module is to explicitly incorporate domain knowledge derived from competition rules into the visual feature aggregation process, thereby providing semantic guidance for rule violation detection. The operational mechanism of the rule-guided spatiotemporal graph convolutional network is illustrated in Figure 1. Human body keypoint coordinates of all players and the soccer ball location within each frame are first extracted using a pretrained pose estimation network and an object detection network, respectively, and are subsequently used as the fundamental inputs for rule quantification. The parameters of the pretrained models are kept fixed throughout the entire training process. Competition rules are then decomposed into a triplet representation consisting of visual preconditions, threshold conditions, and decision outcomes. Consequently, officiating criteria originally expressed in natural language are transformed into frame-wise computable semantic propositions. A Boolean rule trigger is defined as:

$$I_{k,t}(i,j)=1\left[f_{visual}(p_{i,t}^{(j)},p_{i,t}^{(l)},\mathrm{q}_t)\geq\theta_k\right] \tag{1}$$

where, k denotes the rule index, t denotes the frame index, and i and j denote the player-pair indices. The function f_{visual} represents the geometric computation function, which takes the player keypoint coordinates p and the soccer ball position q_t as inputs and produces quantitative measurements, including spatial distance, joint angle, and relative position. The parameter θ_k denotes the decision threshold associated with the k-th rule, and 1[ ] denotes the indicator function, which returns 1 when the specified condition is satisfied and 0 otherwise. The geometric computation function can accommodate multiple rule criteria, including inter-player distance, limb contact distance, leg swing amplitude, and ball possession determination. In this manner, unstructured competition rule text is converted into structured binary trigger signals, thereby providing computable semantic constraints for subsequent feature aggregation.

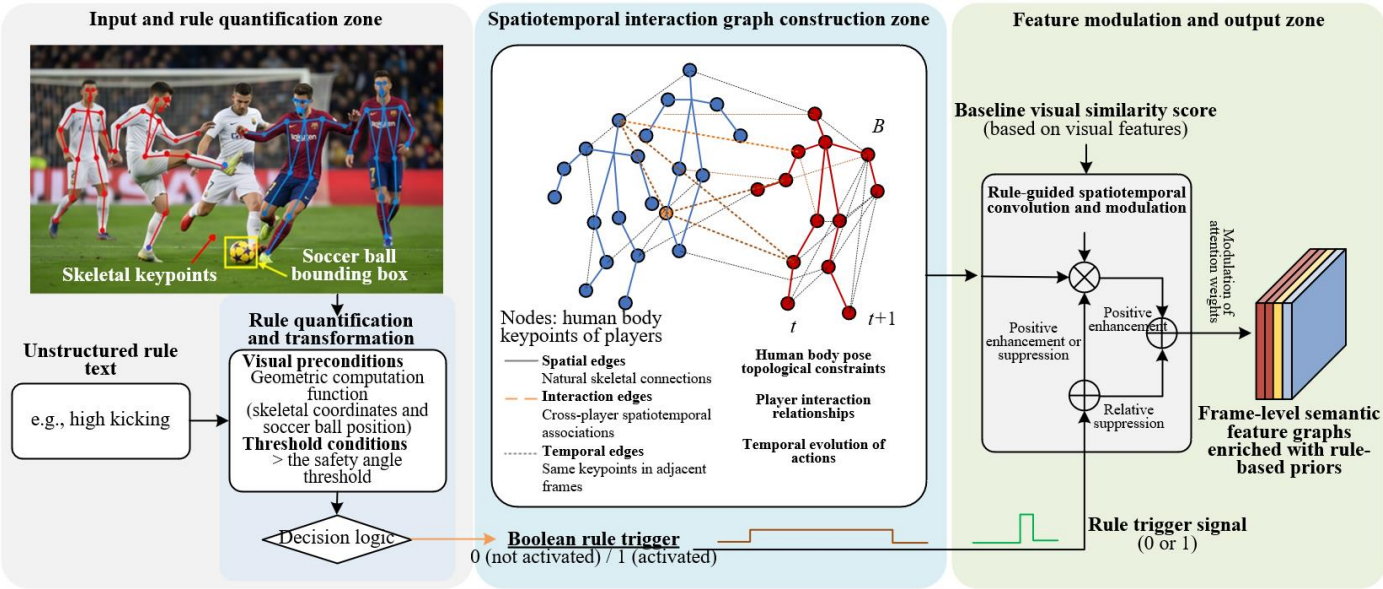

Figure 1. Operational mechanism of the rule-guided spatiotemporal graph convolutional network

A multi-player spatiotemporal interaction graph is subsequently constructed based on the rule trigger signals. Human body keypoints of all players within each frame are treated as graph nodes, whereas the edge set is partitioned into three categories. Spatial edges correspond to the natural skeletal connections of an individual player and encode topological constraints of human body pose. Interaction edges connect keypoint pairs belonging to different players whose spatial distance is smaller than a predefined threshold and encode inter-player interaction relationships. Temporal edges connect the corresponding keypoints across adjacent frames and encode the temporal evolution of actions. The initial node representations are obtained by concatenating the local visual representations extracted by the backbone network with the corresponding keypoint coordinate features. Unlike the standard graph attention network, in which attention weights are computed solely from visual representations, a rule constraint term is incorporated into the attention computation in the proposed module, thereby enabling domain knowledge to directly modulate information propagation. The node feature update rule for the l-th graph convolutional layer is formulated as:

$$h_v^{(l+1)}=\sigma\left(W_0^{(l)}h_v^{(l)}+\sum_{u\in N(v)}\alpha_{vu}^{(l)}\cdot W_1^{(l)}h_u^{(l)}\right) \tag{2}$$

where, $h_v^{(l)}$ denotes the feature vector of node v at the l-th layer; $N(v)$ denotes the spatiotemporal neighborhood of node v; $W_0^{(l)}$ and $W_1^{(l)}$ denote the learnable weight matrices; σ denotes the nonlinear activation function; and $\alpha_{vu}^{(l)}$ denotes the attention weight assigned to neighboring node u with respect to node v.

The attention weights are jointly determined by visual similarity and rule constraints and are computed as:

$$\alpha_{vu}^{(l)}=\frac{exp\left(\beta_{vu}^{(l)}+\gamma\sum_{k=1}^{K}\lambda_k I_{k,t}(v,u)\right)}{\sum_{w\in N(v)}exp\left(\beta_{vw}^{(l)}+\gamma\sum_{k=1}^{K}\lambda_k I_{k,t}(v,w)\right)} \tag{3}$$

where, $\beta_{vu}^{(l)}$ denotes the baseline attention score computed from the original visual representations and reflects the feature similarity between nodes. The parameter γ denotes the global scaling coefficient for rule constraints, whereas λ_k denotes the weight associated with the k-th competition rule. Both parameters are automatically optimized during model training. The term $I_{k,t}(v,u)$ represents the rule trigger associated with the corresponding node pair. When the interaction between two nodes satisfies the rule activation condition, the attention weight of the corresponding edge is positively enhanced, thereby increasing the contribution of the associated interaction features during feature aggregation. Neighborhoods in which no rule is activated retain their baseline attention weights, resulting in a relative suppression effect. Through this mechanism, the feature aggregation process is guided to focus preferentially on interaction patterns that are semantically consistent with competition rules, thereby effectively improving both feature discriminability and semantic specificity. Following multiple layers of spatiotemporal graph convolution and global average pooling, a frame-level semantic representation enriched with rule-based prior knowledge, denoted as F_t^{rule}, is generated and subsequently propagated to the downstream modules to support foul classification and severity assessment.

2.3 Dynamically weighted dual-branch spatiotemporal query fusion network

To satisfy the requirements of spatiotemporal feature aggregation for multi-view match videos, spatiotemporal feature encoding is first performed independently for each view. A video vision Transformer network employing factorized spatiotemporal attention is adopted as the backbone for single-view feature extraction. Synchronized video sequences from each camera view are encoded separately to obtain the initial spatiotemporal feature sequences. To preserve both the spatial positional attributes and the identity information associated with each view, a learnable view embedding vector is added to the feature representation of each view, yielding the view-aware feature representation:

$$Z_v^{(view)}=Z_v+e_v^{view} \tag{4}$$

where, Z_v denotes the original spatiotemporal feature representation extracted from the v-th view, and $Z_v^{(view)}$ denotes the learnable embedding vector corresponding to the v-th camera view. Conventional multi-view fusion methods have generally relied on static fusion strategies, such as direct feature concatenation or fixed-weight feature aggregation. Consequently, fusion weights cannot be adaptively adjusted according to the occlusion status of the action region or the visual quality of individual viewpoints. In densely contested match scenarios, the fusion process is therefore susceptible to interference from low-quality views, thereby limiting further improvements in detection accuracy. The architecture of the dynamically weighted dual-branch spatiotemporal query fusion network is illustrated in Figure 2.

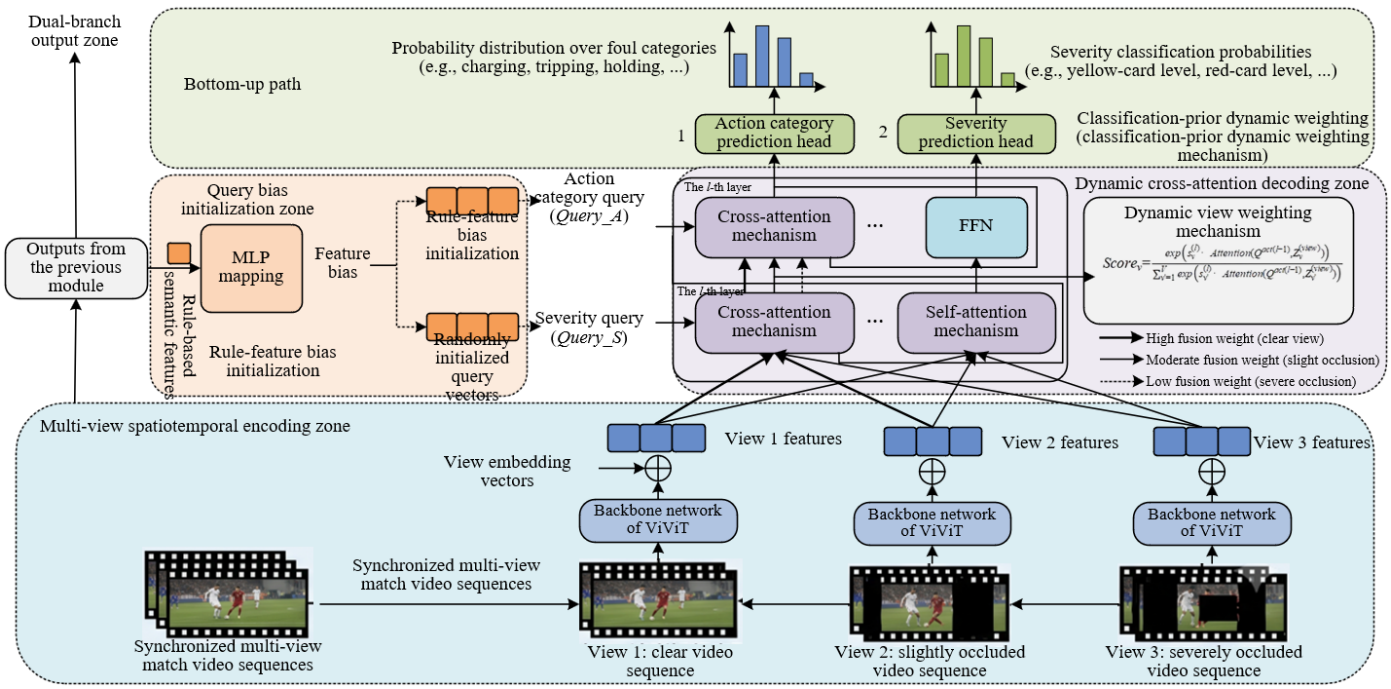

Figure 2. Architecture of the dynamically weighted dual-branch spatiotemporal query fusion network

To emulate the decision-making process by which human referees verify events from multiple viewpoints and preferentially select the most informative observations, a dual-branch spatiotemporal query decoding architecture is constructed. Two categories of learnable query vectors, corresponding to action category and severity assessment, are designed to address the two subtasks of foul classification and severity evaluation, respectively. Adaptive aggregation of multi-view features is achieved through a cross-attention mechanism. During the decoding process, a dynamic view weighting mechanism is introduced. At each decoding layer, the fusion weight associated with each view is dynamically

computed according to the matching degree between the current query state and the corresponding view representation:

$$Score_v = \frac{exp\left(s_v^{(l)} \cdot Attention(Q^{act(l-1)}, Z_v^{(view)})\right)}{\sum_{v'=1}^{V} exp\left(s_{v'}^{(l)} \cdot Attention(Q^{act(l-1)}, Z_{v'}^{(view)})\right)} \quad (5)$$

where, l denotes the number of the decoder layers, V denotes the total number of camera views, $Q^{act(l-1)}$ denotes the action query vector generated by the previous layer, $Attention(\cdot)$ denotes the cross-attention matching score between the query and the view representation, and $s_v^{(l)}$ denotes the learnable confidence factor associated with the v-th view at the l-th layer. After Softmax normalization, $Score_v$ represents the fusion weight assigned to the v-th view during the current decoding step. When the target action region is occluded in a particular view, the matching score between the query and the corresponding view representation is reduced, causing the associated fusion weight to decrease automatically. Conversely, views providing unobstructed observations receive correspondingly higher weights, thereby enabling adaptive adjustments of fusion weights. Following multiple decoding iterations, preliminary probability distributions of foul category classification and severity classification probabilities are generated.

To achieve semantic continuity between the preceding rule-guided module and the current module, a feature bias coupling strategy is adopted to incorporate rule-constrained semantic information into the initial decoder state. Unlike strongly coupled approaches that directly replace the query representations, the proposed design incorporates rule-based representations as a bias term added to the learnable query vectors. The initialization of the action query is formulated as:

$$Q^{act(0)} = Q^{act}_{learnable} + MLP(f_t^{rule}) \quad (6)$$

where, $Q^{act}_{learnable}$ denotes the randomly initialized learnable action query vector, and f_t^{rule} denotes the frame-level semantic representation enriched with rule-based constraints generated by the preceding module. The latter is projected into the same feature space as the query through a multilayer perceptron. Through this design, the initial decoder state is endowed with prior semantic knowledge derived from competition rules, thereby guiding the attention mechanism to focus more rapidly on player interaction regions that are semantically consistent with rule violations and consequently accelerating model convergence. Furthermore, under few-shot training scenarios, the rule-based prior knowledge provides effective semantic regularization, thereby reducing the dependence of the model on large volumes of annotated data while improving the generalization capability for rare foul categories. The severity query is initialized using the same bias strategy, thereby ensuring that both types of tasks are consistently guided by rule-based prior knowledge.

2.4 Geometric potential field-driven fine-grained assessment and interpretation

To satisfy the requirements for quantitative severity assessment and interpretability of rule-violating actions, multi-scale keypoint refinement is first performed within the candidate action regions identified by the second module. An improved You Only Look Once version 8 Pose (YOLOv8-Pose) network is employed to extract the 17 human body keypoints of the interacting players. A reparameterized multi-scale feature fusion module is incorporated to improve the recall of small-scale limb keypoints. In the reduced candidate regions, high-precision detection is conducted, ensuring positioning accuracy while simultaneously reducing the computational cost. Instantaneous velocity vectors of all keypoints are subsequently computed through frame-to-frame differencing of keypoint coordinates across consecutive frames. These velocity vectors are then combined with the corresponding keypoint detection confidence scores to generate weighted velocity estimates, which serve as the fundamental kinematic data for the subsequent calculation of the geometric potential field. The construction process of the geometric potential field and the closed-loop calibration procedure are illustrated in Figure 3.

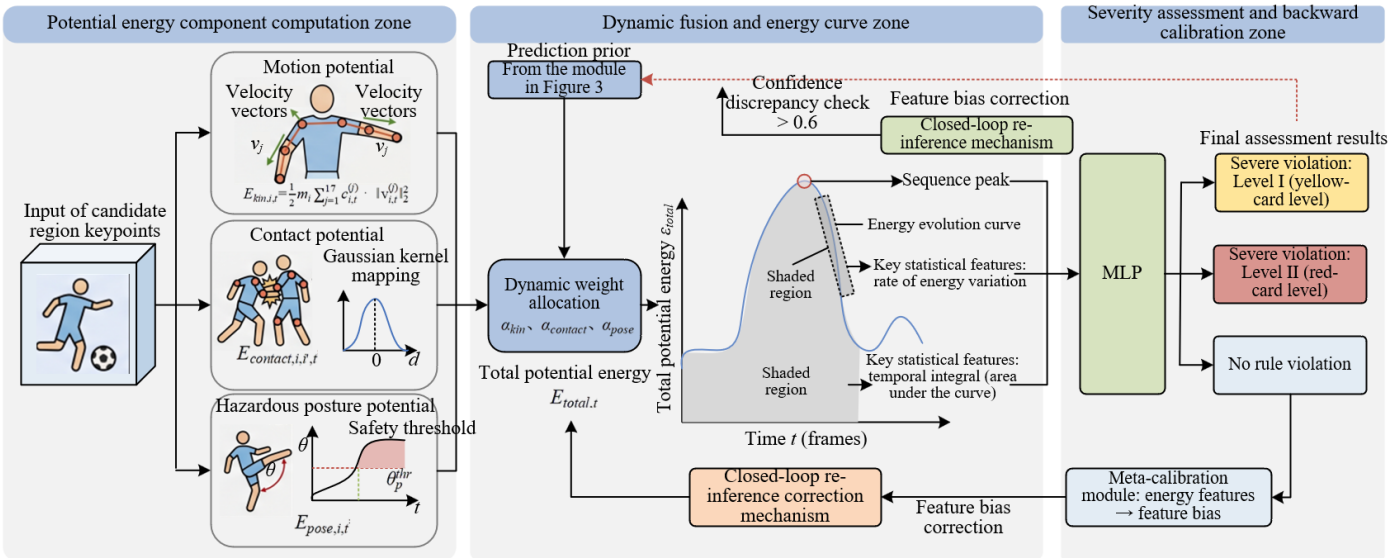

Figure 3. Workflow of geometric potential field construction and the closed-loop calibration

To transform the abstract notion of foul hazardousness into quantifiable and interpretable physical indicators, a geometric potential field model consisting of three potential energy components is constructed. The model characterizes the

severity of rule violations from three complementary perspectives: action intensity, physical contact intensity, and hazardous posture severity. The motion potential is designed to characterize the overall intensity of a player's movement and is computed as a weighted aggregation of the velocities of all detected keypoints. Higher values indicate larger motion amplitudes and stronger physical impacts. The motion potential is defined as:

$$E_{kin,i,t}=\frac{1}{2}m_i\sum_{j=1}^{17}c_{i,t}^{(j)}\cdot\|v_{i,t}^{(j)}\|_2^2 \tag{7}$$

where, m_i denotes the equivalent mass coefficient of player i; $c^{(j)}_{i,t}$ denotes the detection confidence of the j-th keypoint; and $v^{(j)}_{i,t}$ denotes the instantaneous velocity vector of the corresponding keypoint. The contact potential characterizes the intensity of physical contact between two players. It is computed by mapping the spatial distances between corresponding keypoint pairs through a Gaussian kernel, while joint-pair weighting coefficients are introduced to distinguish the hazardousness associated with different contact locations. The contact potential is formulated as:

$$E_{contact,i,i',t}=\sum_{j=1}^{17}\sum_{j'=1}^{17}c_{i,t}^{(j)}c_{i',t}^{(j')}\cdot w_{jj'}\cdot exp\left(-\frac{\|k_{i,t}^{(j)}-k_{i',t}^{(j')}\|_2^2}{2\sigma_{contact}^2}\right) \tag{8}$$

where, $k^{(j)}_{i,t}$ denotes the coordinates of the j-th keypoint of player i; $w_{jj'}$ denotes the hazardousness weighting coefficient associated with the corresponding joint pair; and $\sigma_{contact}$ denotes the Gaussian bandwidth governing the contact distance. The hazardous posture potential characterizes the degree of postural rule violation. It is computed from the deviation between critical joint angles and their corresponding reference values under normal postures. Potential energy is accumulated only when the angular deviation exceeds a predefined safety threshold. The hazardous posture potential is expressed as

$$E_{pose,i,t}=\sum_{p\in P}\phi_p\cdot\max\left(0,\frac{|\theta_p(k_{i,t})-\theta_p^{norm}|}{\theta_p^{thr}}-1\right) \tag{9}$$

where, P denotes the set of critical joints, $\theta_p(k_{i,t})$ denotes the instantaneous angle of the p-th joint, θ_p^{norm} denotes the corresponding joint reference value under a normal posture, θ_p^{thr} denotes the predefined safety threshold, and ϕ_p denotes the weighting coefficient associated with the joint.

The three potential energy components are subsequently fused to obtain the frame-level total potential energy, thereby forming an energy evolution sequence spanning the entire action. The weighting coefficients of the three energy components are not treated as fixed constants. Instead, they are dynamically generated from the severity query vector produced by the second module through a lightweight mapping network. Consequently, different foul categories are associated with distinct energy-component weight allocations. For example, the weight assigned to the contact potential is increased for charging fouls, whereas the contribution of the hazardous posture potential is emphasized for tripping fouls. In this manner, high-level classification priors are allowed to adaptively guide the underlying quantitative assessment process. The total potential energy is computed as:

$$\begin{aligned}E_{total,t}=&\alpha_{kin}\sum_i E_{kin,i,t}+\alpha_{contact}\sum_{i<i'}E_{contact,i,i',t}\\&+\alpha_{pose}\sum_i E_{pose,i,t}\end{aligned} \tag{10}$$

where, α_{kin} , $\alpha_{contact}$, and α_{pose} denote the dynamically generated weighting coefficients for the motion potential, contact potential, and hazardous posture potential, respectively. Three key statistical features are extracted from the continuous energy sequence. The sequence peak reflects the intensity of the most hazardous moment during the action, the energy variation rate at the peak characterizes the abruptness of the movement, and the temporal integral represents the cumulative influence of the rule-violating action throughout its duration. These three features are subsequently fed into a multilayer perceptron, and the final severity assessment is obtained as:

$$s=\arg\max_{s'}\left[MLP_{sev}\left(\max_t E_{total,t}\,,\ \frac{dE_{total}}{dt}|_{max},\ \int E_{total,t}\,dt\right)\right] \tag{11}$$

Through this mechanism, not only is a discrete severity level produced, but the complete frame-level energy curve is also retained as visual evidence for officiating decisions, thereby achieving a unified framework for quantitative severity assessment and decision interpretability.

The proposed module and the preceding classification module are further integrated through a bidirectional calibration mechanism, forming a closed-loop collaborative framework. Along the forward propagation pathway, the foul category and the preliminary severity assessment predicted by the second module are utilized to dynamically adjust both the weighting coefficients of the total potential energy components and the hazardousness coefficients associated with individual joints. Consequently, the quantitative assessment process is adaptively aligned with the officiating criteria corresponding to the current action type, thereby improving the accuracy of severity assessment. Along the backward calibration pathway, a classification–energy consistency verification criterion is established. When a significant confidence discrepancy is detected between the statistical characteristics of the energy evolution sequence and the classification results, the meta-calibration module is activated to inversely map the energy features into feature bias representations. These feature biases are subsequently injected into the decoder layers of the second module to refine the query vectors and the attention distribution, after which the classification inference is performed again. Through this bidirectional feedback mechanism, local misclassifications produced by individual modules can be effectively corrected, thereby improving both the consistency and the robustness of detection results across diverse scenarios. As a result, a complete closed-loop rule violation detection framework is established, spanning visual representation extraction, rule-guided classification, and physics-based quantitative assessment.

2.5 Joint training strategy and loss functions

A two-stage joint training strategy is adopted to ensure both efficient model convergence and effective collaboration among the proposed modules. The keypoint detection module is first pretrained on a publicly available human pose

estimation dataset. During the subsequent end-to-end training stage, the parameters of this module are kept fixed, whereas the rule-guided spatiotemporal graph convolutional network, the dual-branch spatiotemporal query fusion network, and the geometric potential field-based assessment module are jointly fine-tuned in an end-to-end manner. Through this strategy, semantic consistency across the feature spaces of different modules is maintained while collaborative optimization is facilitated. The overall training objective is formulated as a weighted combination of four loss terms:

$$L_{total}=L_{rule}+L_{query}+\lambda_{pose}L_{pose}+\lambda_{energy}L_{energy} \quad (12)$$

where, L_{rule} is implemented using the focal loss formulation and is applied to the interaction-node classification task in the rule-guided module. By introducing the modulation factor, the loss contribution of the abundant normal interaction samples is reduced, enabling the model to focus preferentially on the relatively few challenging rule-violation samples and thereby accommodating the severe class imbalance characteristic of soccer match scenarios, in which rule-violating actions constitute only a small proportion of all interactions. The loss term L_{query} consists of two components. The primary component is the cross-entropy loss for action classification and severity assessment, which is employed to optimize the classification accuracy of the dual-branch decoder. An additional L1 sparsity regularization term is imposed on the view fusion weights to constrain their distribution, encouraging the model to preferentially select a limited number of informative viewpoints while suppressing the contribution of low-quality views. The loss term L_{pose} is formulated as an object keypoint similarity loss and is used to fine-tune keypoint refinement within the candidate regions. This formulation accommodates pose estimation for players at different spatial scales while ensuring the input accuracy of the potential field computation. Finally, L_{energy} is formulated as an ordinal regression loss based on the cumulative link model and is applied to the severity assessment task. By explicitly exploiting the ordinal relationship among severity levels, this loss function more accurately reflects the hierarchical nature of officiating decisions than conventional classification losses.

All experiments were conducted using a single NVIDIA A100 graphics processing unit under the PyTorch deep learning framework. The AdamW optimizer was employed with a weight decay coefficient of 1×10^{-4}. The initial learning rate was set to 3×10^{-5}, and a linear warm-up strategy followed by cosine annealing scheduling was adopted. The model was trained for 30 epochs with a batch size of 8. The loss balancing coefficients λ_{pose} and λ_{energy} were set to 0.5 and 0.8, respectively, to balance the magnitudes of different task-specific losses. This configuration prevented any individual task from dominating the gradient update process, thereby ensuring the stability of multi-task training. During training, gradient clipping was simultaneously applied by constraining the gradient norm to a maximum of 1.0, thereby suppressing potential gradient fluctuations introduced by the closed-loop feedback pathways. During inference, the backward calibration module employed a confidence discrepancy threshold of 0.6. Re-inference was triggered only when the confidence discrepancy between the classification results and the energy-based quantitative assessment exceeded this threshold. Consequently, the consistency of the final decision was maintained while the additional computational overhead introduced by the re-inference process was effectively controlled.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the overall performance of the proposed rule violation detection framework is systematically evaluated through five groups of progressively designed experiments. The evaluation sequentially covers five aspects: overall performance comparison, effectiveness of the core components, few-shot and cross-domain generalization capability, officiating interpretability and decision consistency, and inference efficiency and engineering feasibility. To ensure the statistical reliability of the experimental results, each experiment was independently repeated three times, and the average performance was reported.

3.1 Experimental setup

Performance evaluation was conducted using two datasets. The primary benchmark dataset was the publicly available SoccerNet-Foul dataset, which contains 1,200 professional soccer match clips covering six common foul categories and normal player interactions. Following the official data partition, the dataset was divided into training, validation, and test sets and was used to benchmark the proposed framework under the single-view setting. To evaluate the effectiveness of the proposed multi-view feature fusion strategy, a self-constructed multi-view soccer rule violation dataset was additionally established. Three synchronized acquisition devices were employed to simulate the broadcast view, touchline view, and penalty-area view commonly used in professional match broadcasting. A total of 860 valid video clips were collected and annotated with eight foul categories and four severity levels. All annotations were independently verified by three nationally certified referees through cross-validation. The dataset was partitioned into training, validation, and test sets using a ratio of 7:1:2.

All experiments were implemented using the PyTorch deep learning framework and were conducted on a single NVIDIA A100 graphics processing unit. The AdamW optimizer was adopted with a weight decay coefficient of 1×10^{-4}. The initial learning rate was set to 3×10^{-5}, and a linear warm-up strategy followed by cosine annealing scheduling was employed. The model was trained for 30 epochs with a batch size of 8. The loss balancing coefficients were set to 0.5 and 0.8, respectively, and the gradient clipping norm was fixed at 1.0. These settings were kept identical to those described in the Methodology section to ensure experimental consistency.

3.2 Comparison with state-of-the-art methods

The overall performance of the proposed framework in foul detection and severity assessment was evaluated by comparing it with representative state-of-the-art methods under both single-view and multi-view settings. The primary quantitative results are summarized in Table 1.

As shown in Table 1, the proposed framework achieved the best performance across all evaluation metrics on both datasets. Under the single-view setting, a frame-level mean average precision of 81.2% was achieved, representing an improvement of 5.8 percentage points over the strongest baseline, FoulDetect. Improvements of more than 10% were

also observed in both severity assessment accuracy and the weighted Kappa coefficient, indicating that the rule-guided feature aggregation strategy effectively enhanced the discriminative capability of the model for fine-grained rule-violating actions. Under the multi-view setting, an even greater performance advantage was observed. The proposed framework outperformed the strongest baseline by 6.5 percentage points in frame-level mean average precision. This improvement can primarily be attributed to the dynamic view weighting mechanism, through which fusion weights are adaptively adjusted according to the degree of occlusion, thereby reducing the influence of low-quality viewpoints. Consequently, the proposed fusion strategy more closely resembles the multi-view verification process employed by professional referees. A more detailed analysis of challenging samples further demonstrated that, in scenarios involving dense player occlusions and high-speed actions, the performance degradation exhibited by the proposed framework was substantially smaller than that observed for the baseline methods. These results demonstrate the robustness of the proposed framework under complex competitive match conditions.

3.3 Validation of core modules

The contribution of each core component was quantitatively evaluated through a progressive ablation study, thereby validating the effectiveness of the proposed design. The experimental results are presented in Table 2. A video vision Transformer backbone with a classification head was adopted as the baseline model. The following components were then incorporated sequentially: a standard graph convolutional network, rule-guided constraint, fixed-weight multi-view fusion, the dynamically weighted dual-branch query, the geometric potential field-based assessment, and finally the closed-loop feedback mechanism.

Table 1. Performance comparison of different methods

Method	Single-View (SoccerNet-Foul)				Multi-View (Self-Constructed Dataset)			
	Frame-Level MAP	Category F1	Severity Accuracy	Weighted Kappa	Frame-Level MAP	Category F1	Severity Accuracy	Weighted Kappa
SlowFast	62.3	60.7	52.4	0.412	65.8	63.5	54.1	0.437
ViViT	68.5	66.9	57.2	0.485	72.1	70.3	59.6	0.513
ST-GCN	65.7	64.1	55.3	0.458	68.9	67.2	57.8	0.481
AGCN	67.2	65.8	56.8	0.476	70.5	69.1	59.2	0.504
SoccerNet-Baseline	70.1	68.5	59.7	0.509	73.4	71.8	61.5	0.532
FoulDetect	75.4	73.8	64.2	0.587	79.2	77.6	66.8	0.615
Proposed Method	81.2	79.6	72.5	0.693	85.7	84.1	78.3	0.762

Note: Mean Average Precision = MAP; Video Vision Transformer = WWT; Spatial Temporal Graph Convolutional Network = ST-GCN; Adaptive Graph Convolutional Network = AGCN.

Table 2. Ablation study of the core modules

No.	Experimental Configuration	Frame-Level Mean Average Precision	Severity Accuracy	Weighted Kappa
1	Baseline (Video Vision Transformer + Classification Head)	68.5	57.2	0.485
2	Baseline + Standard Graph Convolutional Network	70.2	59.1	0.507
3	Baseline + Rule-Guided Graph Convolutional Network	74.8	63.5	0.574
4	Previous Configuration + Fixed-Weight Multi-View Fusion	78.3	65.7	0.602
5	Previous Configuration + Dynamically Weighted Dual-Branch Query	81.5	69.4	0.658
6	Previous Configuration + Geometric Potential Field-Based Assessment	82.7	75.1	0.726
7	Full Model (Including the Closed-Loop Feedback Mechanism)	83.9	78.3	0.762

The ablation results demonstrate that each core component contributes positively to the overall performance. The incorporation of the standard graph convolutional network resulted in only a modest improvement, whereas the introduction of rule-guided graph convolution increased the frame-level mean average precision by 4.6 percentage points, confirming that the explicit incorporation of domain knowledge effectively enhances feature aggregation. When the fixed-weight multi-view fusion strategy was replaced by the dynamically weighted dual-branch structure, the frame-level mean average precision increased by 3.2 percentage points, indicating that adaptive view selection substantially improves the utilization efficiency of multi-view representations. Following the incorporation of the geometric potential field-based assessment, substantial improvements were observed in both severity assessment accuracy and the weighted Kappa coefficient, while a further increase in frame-level mean average precision was also achieved. These results indicate that the introduction of physics-based quantitative features provides complementary information for the classification process, thereby enhancing the fine-grained capability of severity assessment. After the closed-loop feedback mechanism was further incorporated, additional improvements were consistently observed across all evaluation metrics. In particular, the weighted Kappa coefficient increased by 3.6 percentage points, demonstrating that the proposed bidirectional calibration mechanism effectively improves the consistency of officiating decisions.

Figure 4. Demonstration of automatic rule violation detection in soccer match videos

To evaluate the practical applicability and interpretability of the proposed framework in real-world match scenarios, a representative tripping sequence was selected for qualitative analysis. The complete image processing pipeline, spanning from the input video frames to the final rule violation decision, was visualized and analyzed. As illustrated in Figure 4, the proposed framework first stably captured the spatial locations of the attacking player, the defending player, and the ball across consecutive video frames. Continuous tracking of these key entities was maintained through bounding-box localization before contact, at the moment of contact, and after the trip, demonstrating that consistent object localization could be preserved under challenging conditions, including motion blur, limb occlusion, and crowded player interactions. Furthermore, the human pose estimation results clearly revealed the structural relationship between the defending player's extended leg during the tackle and the obstruction encountered by the attacking player's stride. The skeletal keypoint connections not only identified the contact location between the lower limbs of the two players but also captured the subsequent kinematic consequences, including forward body inclination, upper-limb extension, and loss of gait stability exhibited by the attacking player. These observations indicate that the proposed framework extends beyond appearance-based recognition to achieve fine-grained action structure analysis. The temporal evidence enhancement stage further integrated motion trajectory association, magnified contact region analysis, and magnified loss-of-balance state analysis into a unified evidence chain. Consequently, the rule violation decision was no longer determined from isolated static frames but was instead supported by continuous visual evidence describing the formation of player contact, motion obstruction, and the temporal evolution of postural abnormalities. The analyzed sequence was ultimately classified as tripping with a confidence score of 0.93, and a Level 2 (moderate) severity assessment was assigned. The predicted result was fully consistent with the corresponding ground-truth annotation.

3.4 Few-shot and cross-scenario robustness

The objective of this experiment was to validate the advantages of rule-based knowledge injection for few-shot generalization and cross-scenario robustness. Two groups of experiments were conducted, including a few-shot evaluation and a cross-domain evaluation. In the few-shot experiments, a limited number of annotated samples were randomly selected from the training set for fine-tuning, and the performance degradation of the proposed framework was compared with that of the purely data-driven video vision Transformer baseline. The corresponding results are presented in Figure 5.

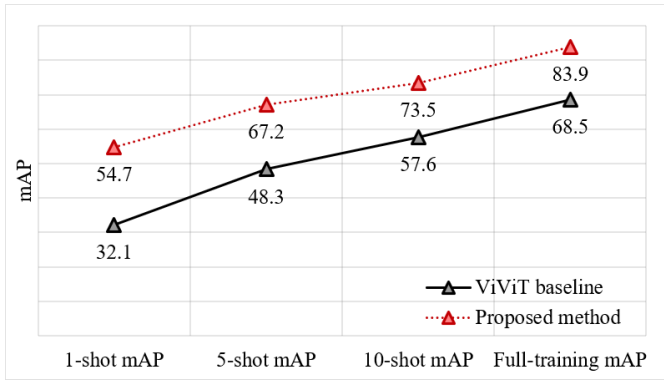

Figure 5. Comparison of generalization performance under few-shot scenarios

As shown in Figure 5, the performance of the purely data-driven approach deteriorated substantially under few-shot learning scenarios. Under the 1-shot setting, the achieved mean average precision was less than half of that obtained with full training. In contrast, the proposed framework maintained a mean average precision of 54.7 under the same 1-shot setting, substantially outperforming the baseline. Moreover, under the 5-shot setting, its performance approached that achieved by the fully trained baseline. These results indicate that the explicit incorporation of competition rules provides

strong semantic priors, thereby effectively reducing the dependence of the model on large quantities of annotated data and yielding superior generalization capability in few-shot scenarios involving rare foul categories. For the cross-domain evaluation, the model trained on the SoccerNet dataset was directly evaluated on cross-league and cross-venue test subsets, thereby assessing its domain generalization capability. The corresponding results are presented in Figure 6. The cross-league subset consisted of match recordings collected from different competition systems, whereas the cross-venue subset contained match clips acquired under different field illumination conditions and camera viewpoints.

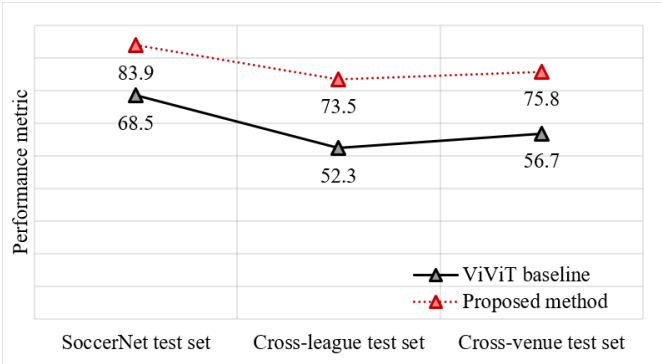

Figure 6. Comparison of cross-scenario generalization performance

The cross-domain evaluation results indicate that the performance of the baseline method deteriorated by more than 20% under cross-scenario conditions, whereas the performance degradation of the proposed framework was limited to less than 12%. Because competition rules possess cross-scenario semantic consistency, the adverse effects of domain shift can be effectively mitigated through rule-based knowledge injection, thereby improving the robustness of the proposed framework across different competitions and diverse data acquisition environments.

3.5 Analysis of interpretability and officiating consistency

The interpretability of the proposed framework and the consistency of its officiating decisions were evaluated in this experiment. The consistency between the model predictions and professional officiating decisions was first assessed through an expert referee evaluation. Three nationally certified referees were invited to independently assign severity levels to 200 test clips. The majority-voting result was adopted as the reference ground truth, and the agreement between the predictions generated by different methods and the referee consensus was quantified. The corresponding results are presented in Figure 7.

The results demonstrate that the proposed framework achieved the lowest mean absolute error in severity assessment, while the weighted Kappa coefficient reached 0.762, indicating a high level of agreement with expert officiating decisions and substantially outperforming all baseline methods. These findings indicate that the geometric potential field-based quantitative assessment effectively aligns the predicted severity levels with the officiating criteria adopted by professional referees, thereby improving the professional credibility of the assessment results.

Visualization of the energy evolution curves for representative cases further demonstrated that the peak of the total potential energy precisely coincided with the temporal occurrence of the rule-violating action, while the decomposed curves of the three energy components clearly characterized the distinctive motion patterns associated with different foul categories. For charging fouls, the contact potential accounted for the largest proportion of the total potential energy. For tripping fouls, the hazardous posture potential was dominant. In contrast, motion potential exhibited the most pronounced increase for foul categories involving rapid movements. The energy evolution curve can therefore serve as interpretable visual evidence for officiating decisions, providing a comprehensive representation of the temporal evolution of action intensity throughout the rule-violating event. Consequently, physically interpretable quantitative evidence is provided to support the final officiating decision, thereby substantially enhancing the interpretability of the proposed framework.

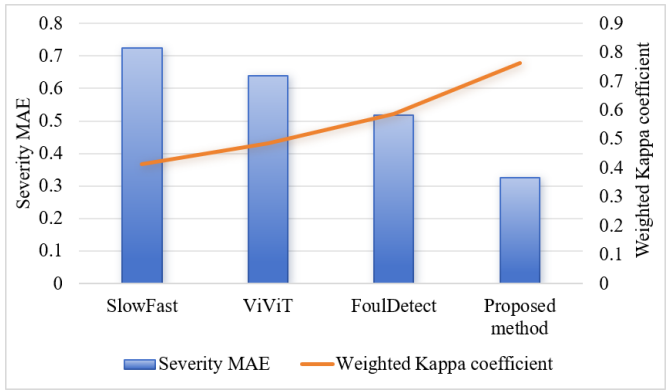

Figure 7. Comparison of officiating consistency and quantitative error

3.6 Analysis of inference efficiency and engineering feasibility

The inference latency and processing speed of the proposed framework were evaluated to assess its engineering feasibility for practical deployment. Table 3 summarizes the latency breakdown of each module in the complete framework together with the inference frame rates achieved on different hardware platforms. All evaluations were conducted using a single-channel video with a resolution of 1080p.

The latency analysis indicates that feature extraction and the dual-branch query decoder constitute the primary computational cost of the proposed framework, whereas the closed-loop calibration introduces an average additional latency of only 2.0 ms, resulting in a negligible impact on the overall inference efficiency. On a mainstream server-grade graphics processing unit, an inference speed exceeding 40 frames per second was achieved, satisfying the requirements for real-time rule violation detection. On a consumer-grade graphics processing unit, near real-time performance was maintained. Furthermore, an inference speed exceeding 10 frames per second was achieved on an edge computing platform, and the processing speed can be further improved through input resolution reduction and model compression. Overall, the proposed framework demonstrates favorable computational efficiency and can be adapted to a variety of deployment scenarios, including live match broadcasting and sideline officiating assistance, thereby demonstrating strong engineering feasibility for practical applications.

Table 3. Analysis of model inference latency and frame rate

Module	Per-Frame Latency (ms)	Percentage	Hardware Platform	Inference Speed (Frames Per Second)
Single-View Feature Extraction	9.7	42.50%	NVIDIA A100	43.8
Rule-Guided Graph Convolution	3.8	16.70%	NVIDIA RTX 3090	28.5
Dual-Branch Query Decoder	5.2	22.80%	Jetson Xavier NX	11.2
Geometric Potential Field Computation	2.1	9.20%	-	-
Closed-Loop Calibration (Average)	2	8.80%	-	-
Total	22.8	100%	-	-

4. DISCUSSION

Based on the comprehensive experimental results, the proposed framework achieved the expected improvements in detection accuracy, officiating consistency, and generalization capability. Nevertheless, three major limitations remain. First, the reliability of the proposed framework remains limited under scenarios involving severe full-body occlusion. When multiple players engage in dense physical interactions and the critical limbs required for rule violation assessment are completely occluded, both the accuracy and the recall of human pose estimation are substantially degraded. The resulting errors are subsequently propagated to the rule triggering and geometric potential field computation stages, thereby increasing the likelihood of missed detections and false predictions. Second, the adaptability of the proposed framework to extremely rare foul categories remains constrained. The current rule formalization strategy primarily covers high-frequency rule-violating actions encountered in competitive matches. For uncommon or highly infrequent foul categories, both the available rule-based priors and the corresponding training samples remain limited. Consequently, a certain degree of degradation in the model's generalization capability is still observed. Third, the depth ambiguity introduced by two-dimensional image projection cannot be completely eliminated. All geometric measurements are currently performed on image-plane coordinates without access to genuine three-dimensional spatial information. As a result, inherent projection errors are introduced into the estimation of player-to-player distances and limb contact distances, thereby limiting the absolute accuracy of both contact potential and hazardous posture assessment.

To address the aforementioned limitations, future research can be directed toward targeted improvements in several key technical aspects. First, multi-view three-dimensional human reconstruction can be integrated to recover the three-dimensional body pose and spatial positions of players through stereo correspondence across synchronized multi-view video streams. By replacing two-dimensional keypoints with three-dimensional pose representations as the input for geometric computation, projection errors can be fundamentally eliminated while the robustness of human pose estimation under severe occlusion can be further improved. Second, audio-modal features can be incorporated to assist contact assessment. A multimodal detection framework can be established by integrating collision sounds, body contact sounds, and other synchronously acquired acoustic signals, thereby compensating for the inherent limitations of purely vision-based approaches in contact determination and further improving the confidence of the detection results.

From a broader perspective, the proposed technical framework possesses considerable potential for extension to a wider range of competitive sports. The rule-based knowledge encoding, dynamic multi-view fusion, and physics-based quantitative interpretation strategies validated in the soccer scenario can be readily adapted to rule violation detection tasks in other multiplayer competitive sports, including basketball, handball, and hockey. Such adaptation can be accomplished by replacing the corresponding competition rule base and the associated quantitative parameter thresholds for each sport. In the long term, the integration of symbolic knowledge guidance with data-driven learning has the potential to facilitate the establishment of a unified intelligent officiating framework applicable to multiple sports, thereby providing sustained technical support for fair officiating and the intelligent advancement of competitive sports.

5. CONCLUSION

To address the technical challenges of insufficient domain knowledge integration, rigid multi-view fusion mechanisms, and the lack of quantitative interpretability in fine-grained rule violation detection for competitive sports, a three-module serial–feedback coupled multi-view soccer rule violation detection framework was proposed, producing foul category classification, severity assessment, and interpretable energy evolution results in an end-to-end manner. Within the proposed framework, competition rules were formalized as computable semantic triggers, enabling explicit guidance of visual feature aggregation through rule-guided spatiotemporal graph convolution. A dynamically weighted dual-branch spatiotemporal query decoder was constructed to adaptively allocate multi-view fusion weights according to the degree of action occlusion, thereby emulating the multi-view verification strategy employed by professional referees. In addition, a geometric potential field model consisting of three components was established to provide physics-based quantitative severity assessment together with interpretable visual explanations. Furthermore, bidirectional collaboration among the three modules was achieved through rule-feature bias initialization, dynamic adjustment based on classification priors, and backward calibration driven by energy inconsistency, thereby establishing a closed-loop optimized detection framework. Experimental results obtained on the publicly available SoccerNet-Foul dataset and the self-constructed multi-view soccer rule violation dataset demonstrated that the proposed framework consistently outperformed representative state-of-the-art methods in terms of frame-level detection accuracy, officiating consistency, and few-shot generalization capability. Moreover, robust detection performance was maintained under complex competitive

match scenarios.

The present study enriches the technical paradigm for the deep integration of symbolic knowledge and visual representations, providing a new solution for fine-grained behavior detection under strong rule-constrained scenarios while offering a practically deployable framework for intelligent image analysis in competitive sports. With appropriate adaptation, the proposed framework can be extended to other multiplayer competitive sports, including basketball and handball, and can also be applied to related tasks such as action quality assessment and intelligent training analysis, thereby demonstrating broad potential for both future academic research and practical engineering applications.

REFERENCES

[1] Zhou, L., Gao, Y., Zhang, M., Wu, P., Wang, P., Zhang, Y. (2024). Human-centric behavior description in videos: New benchmark and model. IEEE Transactions on Multimedia, 26: 10867-10878. https://doi.org/10.1109/TMM.2024.3414263

[2] Vahdani, E., Tian, Y. (2023). Deep learning-based action detection in untrimmed videos: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4302-4320. https://doi.org/10.1109/TPAMI.2022.3193611

[3] Wang, B., Zhao, Y., Yang, L., Long, T., Li, X. (2024). Temporal action localization in the deep learning era: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(4): 2171-2190. https://doi.org/10.1109/TPAMI.2023.3330794

[4] Naik, B.T., Hashmi, M.F., Bokde, N.D. (2022). A comprehensive review of computer vision in sports: Open issues, future trends and research directions. Applied Sciences, 12(9): 4429. https://doi.org/10.3390/app12094429

[5] Host, K., Ivašić-Kos, M. (2022). An overview of human action recognition in sports based on computer vision. Heliyon, 8(6): e09633. https://doi.org/10.1016/j.heliyon.2022.e09633

[6] Nergård Rongved, O.A., Stige, M., Hicks, S.A., Thambawita, V.L., Midoglu, C., Zouganeli, E., Johansen, D., Riegler, M.A., Halvorsen, P. (2021). Automated event detection and classification in soccer: The potential of using multiple modalities. Machine Learning and Knowledge Extraction, 3(4): 1030-1054. https://doi.org/10.3390/make3040051

[7] Cioppa, A., Deliège, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M. (2022). Scaling up SoccerNet with multi-view spatial localization and re-identification. Scientific Data, 9: 355. https://doi.org/10.1038/s41597-022-01469-1

[8] Cioppa, A., Giancola, S., Somers, V., Magera, F., et al. (2024). SoccerNet 2023 challenges results. Sports Engineering, 27(2): 24. https://doi.org/10.1007/s12283-024-00466-4

[9] Luis Del Campo, V., Morenas Martín, J. (2020). Influence of video speeds on visual behavior and decision-making of amateur assistant referees judging offside events. Frontiers in Psychology, 11: 579847. https://doi.org/10.3389/fpsyg.2020.579847

[10] Spitz, J., Wagemans, J., Memmert, D., Williams, A.M., Helsen, W.F. (2021). Video assistant referees (VAR): The impact of technology on decision making in association football referees. Journal of Sports Sciences, 39(2): 147-153. https://doi.org/10.1080/02640414.2020.1809163

[11] Zhang, Y., Li, D., Gómez-Ruano, M.Á., Memmert, D., Li, C., Fu, M. (2022). The effect of the video assistant referee (VAR) on referees' decisions at FIFA Women's World Cups. Frontiers in Psychology, 13: 984367. https://doi.org/10.3389/fpsyg.2022.984367

[12] Samuel, R.D., Galily, Y., Filho, E., Tenenbaum, G. (2020). Implementation of the video assistant referee (VAR) as a career change-event: The Israeli Premier League case study. Frontiers in Psychology, 11: 564855. https://doi.org/10.3389/fpsyg.2020.564855

[13] Mather, G. (2020). A step to VAR: The vision science of offside calls by video assistant referees. Perception, 49(12): 1371-1374. https://doi.org/10.1177/0301006620972006

[14] Zglinski, J. (2022). Rules, standards, and the video assistant referee in football. Sport, Ethics and Philosophy, 16(1): 3-19. https://doi.org/10.1080/17511321.2020.1857823

[15] Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J. (2023). Human action recognition from various data modalities: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 3200-3225. https://doi.org/10.1109/TPAMI.2022.3183112

[16] Shaikh, M.B., Chai, D., Islam, S.M.S., Akhtar, N. (2024). From CNNs to transformers in multimodal human action recognition: A survey. ACM Transactions on Multimedia Computing, Communications, and Applications, 20(8): 260. https://doi.org/10.1145/3664815

[17] Fu, J., Gao, J., Xu, C. (2022). Learning semantic-aware spatial-temporal attention for interpretable action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 32(8): 5213-5224. https://doi.org/10.1109/TCSVT.2021.3137023

[18] Pan, J.H., Gao, J., Zheng, W.S. (2022). Adaptive action assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 8779-8795. https://doi.org/10.1109/TPAMI.2021.3126534

[19] Jain, H., Harit, G., Sharma, A. (2021). Action quality assessment using Siamese network-based deep metric learning. IEEE Transactions on Circuits and Systems for Video Technology, 31(6): 2260-2273. https://doi.org/10.1109/TCSVT.2020.3017727

[20] Du, Z., He, D., Wang, X., Wang, Q. (2024). Learning semantics-guided representations for scoring figure skating. IEEE Transactions on Multimedia, 26: 4987-4997. https://doi.org/10.1109/TMM.2023.3328180