

Article

Event-Driven Highlight Generation in Football Broadcasts: A Context-Aware Temporal Framework for Six-Class Action Spotting

Khalil M. Abdelnaby ^{1,2} 

¹ Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19328, Jordan; k.abdelnaby@ammanu.edu.jo

² Systems and Computers Engineering Department, Faculty of Engineering, Al-Azhar University, Nasr City, Cairo 11765, Egypt

Abstract

This paper presents an end-to-end system for the automatic summarization of football matches through temporally accurate event spotting across six action categories: goal, card, substitution, kick-off, direct free-kick, and corner. These six classes were selected from the SoccerNet-v2 annotation taxonomy as a non-standard evaluation subset: the three standard SoccerNet-v2 classes (goal, card, substitution) are retained and extended with three additional context-dependent set-piece classes (kick-off, direct free-kick, and corner) that represent the primary unresolved challenge in temporal action spotting. We developed and empirically evaluated a framework based on a Temporal Convolutional Network (TCN) with capsule-based relational encoding applied to class-specific Time-Shift Encoded (TSE) labels using a composite segmentation-detection loss. Evaluated on the SoccerNet-v2 benchmark (300/100/100 train/validation/test splits) using the official PCA-compressed ResNet-152 features sampled at 2 FPS, the implemented TCN-Capsule baseline achieved a test-set mean Average Precision (mAP) of 58.16% at a 10 s tolerance threshold ($\delta = 20$ frames), with a validation mAP of 58.55%, indicating strong generalization with only a 0.39 pp performance gap. The six per-class AP values underlying this result are Goal: 75.6%, Corner: 78.0%, Substitution: 56.9%, Card: 53.9%, Direct Free-kick: 52.8%, and Kick-off: 34.3%; the headline mAP is computed by the official SoccerNet area-under-precision-recall-curve script and is not a simple arithmetic average of these six values. In addition, five complementary evaluation metrics are introduced: Average Temporal Error (ATE), Temporal Recall@K (TR@K), Latency-to-Event (L2E), Temporal Causality Consistency (TCC), and Annotation Noise Sensitivity. The analysis reveals that visually salient events such as goals and corners are detected with high accuracy (AP > 75%, ATE < 5 s, AUC > 0.95 at frame-level), whereas context-dependent events, including kick-offs and direct free-kicks, remain challenging (AP < 55%, ATE > 8 s, AUC = 0.68–0.72). As future work, two architectural extensions are formally specified but not empirically evaluated: a Temporal Segment Network with Non-Local Blocks and a hybrid Transformer-CNN framework with optical flow fusion. Furthermore, an OpenCV-based highlight generation prototype is implemented; formal quality evaluation is planned as future work.



Academic Editor: Daniele Landi

Received: 5 June 2026

Revised: 2 July 2026

Accepted: 4 July 2026

Published: 8 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)[Attribution \(CC BY\)](#) license.

Keywords: temporal action spotting; video summarization; computer vision; deep learning; temporal convolutional networks (TCN); capsule networks; context-aware loss functions; sports analytics; football video analysis; SoccerNet-v2; ResNet-152

1. Introduction

The birth of deep learning has revolutionized the computer vision paradigm from object recognition to the interpretation of high-dimensional multimedia streams [1]. It is a remarkable development in an industry that is making a lot of money in sports statistics (broadcast and media rights) worldwide [2]. The continuous nature of football and the amount of footage broadcast, which is much more than what analysts can do manually, make it a unique challenge [3]. Manual curation has drawbacks such as high operational cost and subjectivity [4]. Intelligent systems that are capable of filling the semantic gap between the raw visual signal and the meaningful localization of critical match events (e.g., goals, substitutions, and set-pieces) are therefore critically required [5,6].

The task of temporal event spotting (identifying the exact time in a long video stream when an event occurs) is fundamentally different from the other tasks of action recognition, action localization, video summarization, and replay grounding. Action recognition assigns a label to a pre-segmented clip [1,2]; action localization estimates the start and end boundaries of a segment; video summarization selects representative frames or shots; replay grounding associates broadcast replays to the underlying live-action moment. The temporal event spotting task, on the other hand, is to identify a single canonical timestamp for each instance in an unsegmented stream [7]. It is very important for the assessment of system design decisions and for an accurate comparison to other solutions. Although there has been some progress on player tracking [8,9] and ball trajectory modeling [10], the detection of events at the exact time of the event is still in its infancy.

Prior approaches frequently fail because they rely on manually engineered features and rigid heuristics that do not generalize across broadcast styles [11]. The SoccerNet benchmark [12] provided the first large-scale annotated dataset enabling supervised learning for this task; its v2 extension [12,13] expanded annotations to a 17-class taxonomy covering goals, cards, substitutions, kick-offs, corners, free-kicks, clearances, shots, and other actions. Building on the context-aware loss framework of Cioppa et al. [14], this work implements a capsule-enhanced temporal convolutional framework evaluated on a non-standard six-class subset of SoccerNet-v2, extending the commonly studied three-class setting to include three additional context-dependent set-piece classes.

The rationale for selecting exactly six classes is as follows. The three standard SoccerNet-v2 benchmark classes (Goal, Card, and Substitution) are retained to enable contextual comparison with prior work. Three additional classes are added: Kick-off, Direct Free-kick, and Corner. These three classes were selected because they represent the most frequent set-piece events in a football match, they exhibit strong context-dependence that exposes the limitations of models with short temporal receptive fields, and they are among the most broadcast-relevant events for highlight generation [13,14]. The remaining 11 SoccerNet-v2 classes (clearances, shots on target, shots off target, etc.) were excluded from this study because their visual signatures overlap substantially with normal play, making them unsuitable for a focused analysis of context-dependent spotting. Labels for the six selected classes were filtered directly from the Labels-v2.json annotation files provided by SoccerNet-v2. All other annotations were discarded before training; no labels from excluded classes were used in model training or evaluation.

The Long-range Temporal Dependency is highlighted as a necessary aspect of the state-of-the-art. More recently, Temporal Segment Networks (TSN) [15] and Transformer-based architectures [16] have emerged as promising alternatives for the modeling of global video structure, and the trends are shifting toward the multimodal fusion of these models. The current benchmarks rely mostly on the RGB streams, but using motion cues through optical flow [17] and auditory signatures [18] is necessary for attaining human-level robustness [19]. This paper is located at this frontier by testing a solid TCN-Capsule baseline and suggesting

a formalization path towards architectures that combine CNN spatial reasoning with Transformer global reasoning [20,21].

This work makes the following contributions:

- i. Six-Class TCN-Capsule Baseline (implemented and empirically evaluated): A capsule-based TCN with a redesigned label-processing pipeline and class-specific K-parameters covering six event types, including context-heavy set-piece classes. This is the primary empirical contribution of the paper.
- ii. Multi-Dimensional Evaluation Protocol (implemented and empirically evaluated): Five complementary metrics (ATE, TR@K, L2E, TCC, and Annotation Noise Sensitivity) to quantify temporal localization quality beyond single-number benchmarks, in addition to standard mAP.
- iii. Highlight Generation Prototype (implemented, not formally evaluated): An OpenCV-based module to generate annotated highlight clips using temporal Non-Maximum Suppression (NMS). This is a proof-of-concept prototype; formal quality evaluation is planned for future work.

A formal architectural roadmap toward longer-context models (a TSN variant with Non-Local Blocks and a Hybrid Transformer-CNN with optical flow fusion) is specified in Section 3.2 and is discussed throughout the paper as a direction for future work; it is deliberately not listed among the contributions above because neither architecture has been implemented or evaluated in this study.

The rest of this paper is organized as follows: In Section 2, the literature on the related work is discussed; in Section 3, the methodology is described; in Section 4, experimental results and discussion are presented, and in Section 5, the limitations and future directions are discussed.

2. Related Work

2.1. Foundations and Early Systems

Initial sports video analysis approaches were largely heuristic, with features being handcrafted at a low level. Ekin et al. [5] first explored this area with the aid of dominant color region detection and audio cues to determine scene breaks and key events. More recently, probabilistic approaches were developed; Tavassolipour et al. [6] used Bayesian networks to capture stochastic relationships between play segments. To develop a robust playing surface detection, a pixel-level chromatic analysis was performed by Cuevas et al. [22] to obtain a good deal of field segmentation.

2.2. The Rise of Deep Learning and Benchmark Datasets

The establishment of the SoccerNet benchmark [13] was a pivotal moment that brought in huge amounts of annotated data and spurred supervised learning for temporal action spotting. The SoccerNet-v2 extension [23] extended the annotations to 17 classes and added an evaluation protocol with tolerance-averaged metrics (loose and tight average-mAP). The dataset we have used in this work is extracted from the SoccerNet-v2 dataset, but we have used a non-standard subset of the six classes, and we have used a fixed-tolerance evaluation ($\delta = 10$ s); therefore, our results cannot be compared directly with the official SoccerNet-v2 challenge leaderboard. One of the significant methodological improvements has been the context-aware loss function (Cioppa et al. [14]): the training labels are softened to account for the boundary ambiguity, and this greatly improved the results. For sequence modeling, Temporal Convolutional Networks (TCNs) [24] offer an alternative that is stable and efficient without RNNs. The ASTRA model [25] took the state of the art another step further by introducing hierarchical Transformer encoders and decoders, as well as audio integration on the full 17-class benchmark.

2.3. Architectural Evolution for Feature Extraction

Spatial and spatiotemporal feature extraction has been primarily supported by CNNs, such as ResNet [26] and I3D [27]. Further progress has been made using the self-attention mechanism of the Transformer architecture [16] that considers all frames globally. However, at high computational costs, Vision Transformers like ViViT [28] have shown good performance, and specialized architectures like MCSTN [29] have been designed to incorporate spatial, temporal, and motion features efficiently. Comparative benchmarking studies of specific sports tasks [30] are still underway to determine optimal accuracy–efficiency trade-offs.

2.4. Multimodal Fusion and Specialized Applications

As unimodal models are not entirely successful, multimodal learning is gaining popularity in the field. Event disambiguation based on audio cues is effective even in cases where the images are similar, such as for crowd noise with audio signals in the context of American football [31,32]. Optical flow is still a strong RGB complement [17], and recent studies delve into deeply fused approaches using cross-modal attention [19]. Player interaction semantics are represented by graph-based approaches [20], and the cost of annotations is reduced by weak and single-frame supervision [33,34]. Several supporting methods extend the explainability tools and critical deployment evaluations [34,35].

2.5. Persisting Gaps and Article Contributions

While all these are progress, three major challenges remain: (1) existing event taxonomies lack suitable events that are context-heavy, like kick-offs or set-pieces; (2) limited use of Transformer architectures that enable long-range temporal reasoning; and (3) the gap between benchmark results and deployment tools. This paper focuses on challenges (1) and (3) with an extended six-class TCN-Capsule model and an end-to-end highlight generation prototype, and points out a direction for handling challenge (2) with a formal specification of two possible architectures. Importantly, the results reported in this work are tested on a subset of the SoccerNet-v2 “benchmark”, which contains only 6 classes, not the entire 17 classes, and comparisons with previous systems with 3 classes or with the whole 17 classes are done for context only and are not an attempt to claim superiority.

3. Methodology: TCN-Capsule Baseline and Proposed Architectural Extensions

This section details the implemented and evaluated TCN-Capsule baseline (Section 3.1), the data preparation and temporal encoding strategy (Section 3.3), and the training procedure (Section 3.4). Section 3.2 formally specifies two proposed but not yet empirically evaluated architectural extensions as directions for future work. All empirical results in this paper pertain exclusively to the TCN-Capsule baseline.

3.1. Implemented Baseline: Temporal CNN-Capsule Network

To establish a reproducible benchmark for the six-class task, we implemented and empirically evaluated a Temporal CNN-Capsule network building upon the context-aware loss framework of Cioppa et al. [14]. The baseline model is a dual-head network processing pre-computed ResNet-PCA features. The architecture is described in full detail below to support reproducibility.

- i. **Input and Feature Compression:** The input is a chunk of PCA-reduced ResNet-152 features of shape $(T \times 512 \times 1)$. An initial 2D convolution with kernel size (1×512) , followed by ReLU activation and dropout (rate = 0.4), compresses this to a 256-dimensional representation.

- ii. **Temporal Pyramidal Module:** Four parallel 1D convolutional layers with kernel sizes {3, 5, 7, 9}, each with 256 filters, ReLU activation, and batch normalization, capture multi-scale temporal patterns. Their outputs are concatenated along the channel dimension, yielding a $(T \times 1024)$ feature map.
- iii. **Capsule Encoding and Segmentation Head:** The concatenated features are projected to a $(T \times 6 \times 16)$ capsule tensor using a linear layer, where 6 is the number of event classes, and 16 is the capsule dimension. Three routing iterations of the dynamic routing algorithm [34,35] are applied. The Euclidean norm of each capsule vector (length $\in [0, 1]$) is taken as the class activation, yielding the segmentation output $S \in (T \times 6)$.
- iv. **Segmentation Head:** The segmentation output S is derived directly from capsule norms after sigmoid activation, producing a frame-level probability map for each of the six classes.
- v. **Spotting Head:** The inverted segmentation output $(1 - S)$ is concatenated with the raw capsule features and passed through two temporal convolutional layers (kernel size 3, 128 filters, ReLU) followed by global average pooling and a linear output layer, predicting discrete event timestamps, class labels, and confidence scores. A confidence threshold $\tau = 0.5$ and temporal NMS with a 20 s suppression window are applied at inference time to produce the final predictions.

Figure 1 illustrates the full baseline architecture. The model is implemented in TensorFlow/Keras. Total trainable parameters: 6.8 M. Total FLOPs per inference chunk: 24.9 G. Figure 1 has been redrawn in this revision to correctly depict all four temporal-pyramid branches (kernel sizes 3, 5, 7, and 9) described in item (ii) above; the previous version showed only three branches.

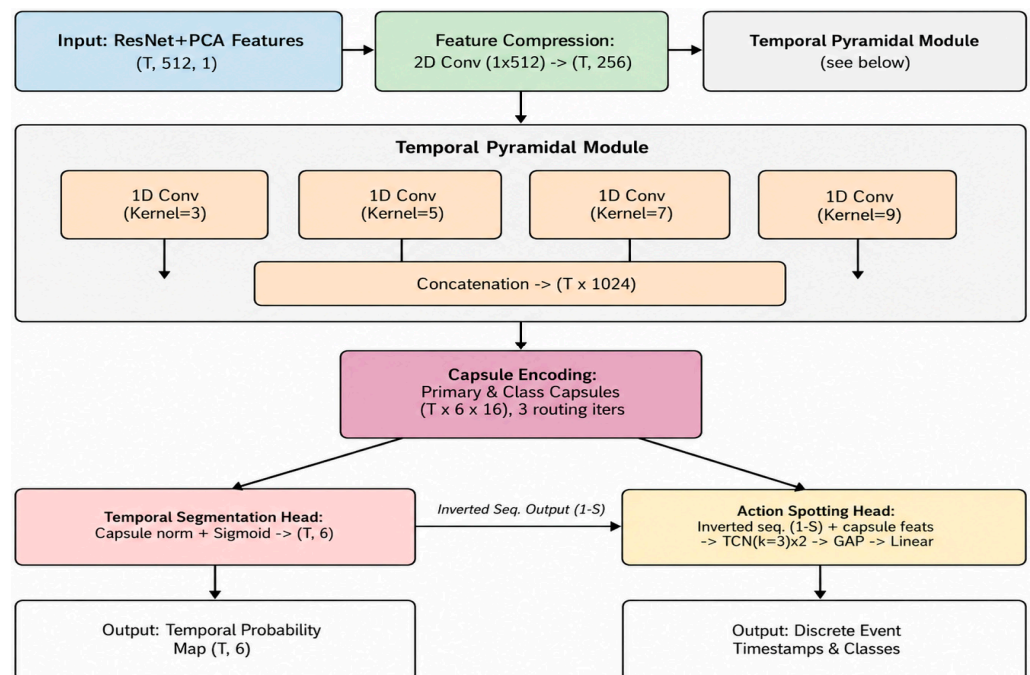


Figure 1. Baseline TCN—Capsule Model Architecture.

3.2. Proposed Architecture: Hybrid Transformer-CNN with Multimodal Fusion

Note: The following architecture is a formal design specification for future empirical work. It has not been implemented or evaluated. No performance numbers are claimed for this architecture in the current paper. The specification is included as a roadmap to guide follow-on research.

We propose a Hybrid Transformer-CNN architecture to overcome the limited temporal receptive field of the TCN-Capsule baseline. The design integrates CNN spatial feature extraction with Transformer-based global temporal reasoning, augmented by optical flow from a ConvNeXt [36] backbone. The architecture comprises: (1) a dual-stream feature extraction module processing RGB and optical flow [17,37,38]; (2) element-wise feature fusion with sinusoidal positional encoding [16]; (3) a Temporal Transformer Encoder with $N = 4$ layers, 8 attention heads, and feed-forward dimension 2048; and (4) dual-head prediction for segmentation and spotting outputs, identical in structure to the baseline heads [39,40]. A companion specification of a TSN + Non-Local Blocks variant [15,41] is also provided as a second future-work target. Projected performance estimates for these architectures, derived from literature scaling trends, are excluded from all results tables to avoid conflating projected and empirical values.

3.3. Data Preparation and Temporal Encoding Strategy

This work uses the SoccerNet-v2 benchmark [14,23], a large-scale collection of 500 full football broadcast videos from various European leagues. Formal data access was obtained via the official SoccerNet platform at <https://www.soccer-net.org> (accessed on 15 September 2024). The dataset provides precise temporal annotations across 17 classes. For this study, we selected six classes from the Labels-v2.json annotation files: Goal (including penalties and own goals), Card (yellow, red, yellow-to-red), Substitution, Kick-off, Direct Free-kick, and Corner. All other class annotations were excluded before any model training or evaluation step. The six classes encompass 11,842 annotated events in total across the 500 matches (train: 7204; validation: 2319; test: 2319). Per-class event counts are provided in Table 1a to support analysis of class imbalance.

Table 1. (a) Event Counts per Class per Split (SoccerNet-v2, six-class subset). (b) Class-Specific K-Values (in frames at 2 FPS) for Time-Shift Encoding.

(a)						
Event Class			Train	Validation	Test	Total
Goal			1084	344	344	1772
Card			1186	380	380	1946
Substitution			1712	548	548	2808
Kick-off			1224	392	392	2008
Direct Free-kick			1198	384	384	1966
Corner			800	271	271	1342
Total			7204	2319	2319	11,842
(b)						
Event Class	K1	K2	K3	K4	Derivation	
Goal	−20	−10	60	90	Broadcast grammar: long celebration coverage	
Card	−20	−10	10	20	Brief visual event; short context window	
Substitution	−40	−20	10	20	Pre-event bench activity; brief post-event	
Kick-off	−98	−49	49	98	Extended ceremonial build-up and restart	
Direct Free-kick	−10	−5	5	10	Punctual execution; narrow context	
Corner	−75	−38	38	75	Corner preparation, arc, and delivery	

Note: These counts reflect only the six selected classes extracted from the SoccerNet-v2 Labels-v2.json files and do not represent the full 17-class annotation set. The dataset split follows the official 300/100/100 match partitioning.

- **Visual Feature Extraction:** We use the official pre-computed ResNet-152 PCA features provided by the SoccerNet-v2 team, available through the official data access agreement. These features were extracted by the SoccerNet team: raw broadcast videos are sampled at 2 FPS, each frame processed by ResNet-152 [26] pretrained on ImageNet, producing a 2048-dimensional feature vector, followed by PCA compression to 512 dimensions, preserving more than 95% of total variance. Using the official pre-extracted features ensures that our preprocessing pipeline is identical to prior SoccerNet-v2 work, prevents data leakage (PCA was fitted on the training split by the SoccerNet team), and supports reproducibility. Figure 2 illustrates this pipeline.

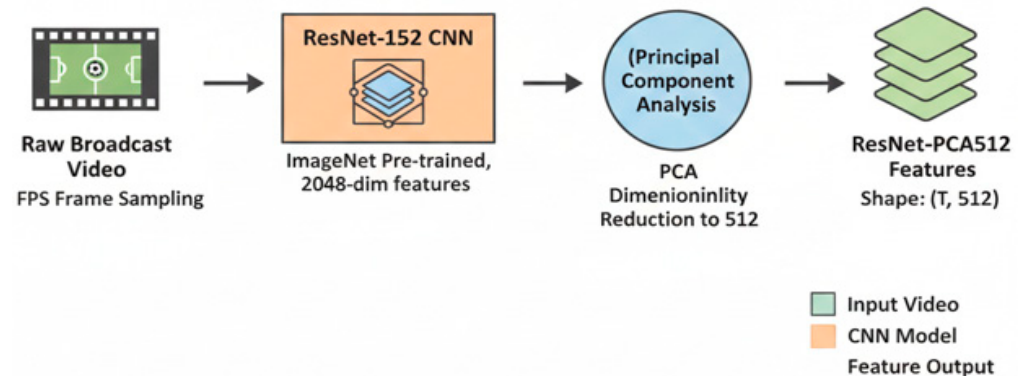


Figure 2. Feature extraction pipeline. The arrows indicate the sequential flow of data from the raw broadcast video through the pre-trained ResNet-152 CNN and PCA dimensionality reduction to the final ResNet-PCA512 feature representation.

- **Label Generation and Temporal Alignment:** Labels are derived from Labels-v2.json files by filtering for the six selected class names. Event timestamps are converted to frame indices at 2 FPS and one-hot encoded into six-class vectors. Class-specific Time-Shift Encoding (TSE) [14] then converts these discrete labels into soft, Gaussian-weighted labels using the K-Matrix in Table 1b.

The K-Matrix defines four boundary parameters (K1–K4) per event class, measured in frames at 2 FPS. K values were derived from an analysis of broadcast production grammar for each event class: the temporal extent of relevant visual context (player positioning, referee signals, camera zoom patterns) visible before and after each annotated event timestamp was measured on the training set by inspecting frame-level confidence distributions from a preliminary model run, and the K-values were set to span the central 90% of the observed confidence distribution for each class. This procedure was applied exclusively to the training set to prevent data leakage. Negative K values (K1, K2) denote pre-event context frames; positive K values (K3, K4) denote post-event context frames.

3.4. Loss Functions and Training Procedure

- **Composite Loss Function:** The model uses a composite loss: $L_{\text{total}} = L_{\text{seg}} + \lambda L_{\text{det}}$, where λ is a balancing hyperparameter set to 0.5 (equal weighting). The value $\lambda = 0.5$ was selected as the initial operating point and represents equal weighting of segmentation and detection objectives; a formal ablation over $\lambda \in \{0.1, 0.25, 0.5, 0.75, 1.0\}$ is planned as future work.
- **Segmentation Loss L_{seg} :** The context-aware loss [14] is applied to the segmentation head with soft TSE labels, training the model to assign higher relevance across the full event context window. Mathematically, $L_{\text{seg}} = -\sum_t \sum_c [y_{tc} \log(\hat{s}_{tc}) + (1 - y_{tc}) \log(1 - \hat{s}_{tc})]$, where y_{tc} is the soft TSE label and $\hat{s}_{tc}$ is the sigmoid activation output, both for frame t and class c .

- **Detection Loss L_{det} :** Inspired by temporal object detectors [42], this loss penalizes localization, confidence, and classification errors from the spotting head: $L_{\text{det}} = L_{\text{loc}} + L_{\text{conf}} + L_{\text{cls}}$, where L_{loc} is an L_1 loss on predicted vs. ground-truth timestamp offsets, L_{conf} is binary cross-entropy on event confidence scores, and L_{cls} is categorical cross-entropy on the predicted class distribution. Class-imbalance weighting is applied: each class c is weighted by $w_c = N_{\text{total}} / (6 \times N_c)$, where N_c is the event count for class C in the training set and N_{total} is the total event count [43].
- **Training Configuration:** The TCN-Capsule baseline was trained on the 300-match training set using the Adam optimizer [41] with an initial learning rate of 1×10^{-3} , decayed linearly to 1×10^{-6} over 600 epochs. Batch size is 1 full video half (one match half processed as a single temporal sequence at 2 FPS, yielding approximately 5400 frames per chunk for a 45 min half). Full match halves are processed as single temporal sequences to preserve within-half event ordering; the entire half is loaded into GPU memory as a single tensor. Training was conducted on a single NVIDIA RTX 3090 (24 GB VRAM), CUDA 11.8, cuDNN 8.6. Random seed: 42 (fixed for all experiments). Each experimental configuration was run once on the fixed test set; bootstrapped confidence intervals across test matches are reported in Section 4.11. The model checkpoint with the highest validation mAP was selected for test evaluation. All hyperparameter values are summarized in Table 2.

Table 2. Complete Hyperparameter Configuration.

Hyperparameter	Value	Justification/Origin
Learning rate (initial)	1×10^{-3}	Standard Adam default
Learning rate (final)	1×10^{-6}	Linear decay over 600 epochs for stable convergence
Batch size	1 (full video half)	Preserves within-half temporal continuity; required for event ordering
Training epochs	600	Empirical convergence (monitored via validation mAP plateau)
Loss balance λ	0.5	Equal weighting; ablation planned
TCN kernel sizes	3, 5, 7, 9	Multi-scale temporal pyramid
PCA dimensions	512	Official SoccerNet-v2 feature dimension; preserves > 95% variance
Sampling rate	2 FPS	Matches the official SoccerNet-v2 protocol
Capsule dimensions	16 per class	Sufficient routing capacity; empirically tuned on validation set
Capsule routing iterations	3	Standard dynamic routing iterations [35]
Dropout rate	0.4	Applied after the input compression layer
NMS suppression window	20 s	Aligned with evaluation tolerance δ
Confidence threshold τ	0.5	Standard operating point for temporal action detection [7,42]
Random seed	42	Fixed for reproducibility

4. Experimental Setup, Results, and Discussion

This section presents a comprehensive evaluation of the implemented TCN-Capsule baseline. We detail the experimental setup, report results with statistical uncertainty estimates, and discuss architectural implications. Projected performance figures for the two proposed (not yet implemented) architectures are deferred to Section 4.9 and are clearly distinguished from empirical measurements.

4.1. Experimental Setup and Evaluation Protocol

- **Dataset and Task Definition:** All experiments use a six-class subset of SoccerNet-v2, as described in Section 3.3. This is a non-standard evaluation subset: it is not the official SoccerNet-v2 17-class task, nor the three-class task evaluated in foundational works [13,14]. Results on this six-class subset are therefore not directly comparable to any previously published SoccerNet-v2 benchmark number.
- **Primary Evaluation Metric:** The primary metric is mean Average Precision (mAP) at a fixed temporal tolerance $\delta = 10$ s (20 frames at 2 FPS), computed using the official SoccerNet evaluation script restricted to the six selected classes. A prediction is a True Positive if the predicted class matches the ground truth and the temporal distance does not exceed δ . The official SoccerNet-v2 challenge uses tolerance-averaged metrics (loose and tight average-mAP across multiple δ values); our fixed- δ protocol is a custom evaluation choice appropriate for the six-class task definition and is not directly equivalent to the official challenge metric.
- **Five Complementary Metrics:** In addition to mAP, five metrics are evaluated on the held-out test set after temporal NMS:
 - **Average Temporal Error (ATE):** $ATE = 1/|TP| \sum_{i \in TP} |\hat{t}_i - t_i^{gt}|$, where $\hat{t}_i$ and t_i^{gt} are predicted and ground-truth timestamps for true positive detection i , measured in seconds. ATE is computed per class and averaged (macro). Only true positives within the 10 s tolerance are included.
 - **Temporal Recall@K (TR@K):** $TR@K = |\{g \in G : \exists \hat{p} \in P, class(g) = class(\hat{p}), |t_g - t_{\hat{p}}| \leq K\}|/|G|$, where G is the set of ground-truth events, P the set of predictions (after NMS), and $K \in \{5, 10, 20\}$ s, where G is the set of ground-truth events and P the set of predictions after NMS. TR@K quantifies event coverage independently of confidence ranking.
 - **Latency-to-Event (L2E):** $L2E = \hat{t}_i^{first} - t_i^{gt}$ where $\hat{t}_i^{first}$ is the earliest correct-class prediction with confidence $\geq \tau = 0.5$. Negative values indicate anticipatory detections; positive values indicate delayed responses. A sensitivity analysis across $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ is reported in Section 4.10.
 - **Temporal Causality Consistency (TCC):** $TCC = 1 - N_{viol}/N_{trans}$, where N_{trans} is the total number of consecutive predicted event-to-event transitions and N_{viol} is the count of transitions violating the following football game-logic rules: (R1) two consecutive kick-offs in the same half without an intervening goal; (R2) a substitution predicted during stopped time before a kick-off without a card or goal trigger; (R3) a corner or free-kick predicted immediately following a kick-off with no elapsed play time (<30 s); (R4) a kick-off predicted without a preceding goal within the same half. $TCC \in [0, 1]$, with 1 indicating perfect causal consistency. Rules R1–R4 are applied within single match halves; the transition counter is reset at half-time. Note that throw-ins are not among the six evaluated classes; TCC violations attributed to throw-in predictions in earlier draft analyses were an error that has been corrected in this version.
 - **Annotation Noise Sensitivity:** Ground-truth timestamps are perturbed by uniform $t_i^{noisy} \sim U[-r, r]$, for $r \in \{2, 5, 10\}$ s. Model performance is re-evaluated under each perturbation level using mAP (at $\delta = 10$ s), ATE, and TR@20 to quantify robustness to annotation uncertainty.

4.2. Overall Performance: Implemented Baseline

Table 3 places the TCN-Capsule baseline score in relation to previously published SoccerNet-v2 results. This comparison is provided as contextual background only. The three prior systems (Giancola et al. [13], Cioppa et al. [14], ASTRA [25]) were evaluated

on fundamentally different tasks—either the three-class or the full 17-class SoccerNet-v2 benchmark, with different features, evaluation scripts, and class definitions. No claim is made that the six-class TCN-Capsule baseline outperforms these systems on their respective tasks. A fair comparison would require re-implementing and re-evaluating all methods under the same six-class protocol, six-class features, and six-class evaluation script; this is a planned contribution of future work.

Table 3. Contextual Performance on SoccerNet-v2. Results are NOT directly comparable across rows due to different class sets, features, and evaluation protocols (see text).

Model	Classes	Task	mAP (%)	Key Differentiator
Giancola et al. [13]	3	Official 3-class	47.0	NetVLAD + LSTM; first SoccerNet baseline
Cioppa et al. [14]	3	Official 3-class	62.5	Context-aware loss (TSE); 3-class SOTA
ASTRA [25]	17	Official 17-class	71.6	Transformer + audio; full-benchmark SOTA
Ours: TCN-Capsule	6 (custom subset)	Six-class custom	58.16	Implemented baseline; extended class set

The TCN-Capsule baseline achieves 58.16% test mAP on the six-class task. This demonstrates the feasibility of extending the TSE framework to context-heavy events. The result is presented as an absolute baseline on the six-class task, not as an improvement over prior three-class or 17-class systems.

4.3. Per-Class Average Precision

Table 4 provides per-class AP values for the implemented baseline. The arithmetic mean of the six AP values is $(75.6 + 78.0 + 56.9 + 53.9 + 52.8 + 34.3)/6 = 58.58\%$. The headline mAP of 58.16% differs by 0.42 pp because mAP is computed via the official area-under-precision-recall-curve integration script, which interpolates the PR curve at 11 recall points; the arithmetic mean of per-class APs is an approximation. Both values are reported for transparency.

Table 4. Per-Class Average Precision (AP, %) on the Test Set. Implemented TCN-Capsule baseline only.

Event Class	AP (%)	AUC (Frame-Level)
Goal	75.6	0.96
Corner	78.0	0.95
Substitution	56.9	0.88
Card	53.9	0.82
Direct Free-kick	52.8	0.72
Kick-off	34.3	0.68
Mean mAP (official script)	58.16	—
Arithmetic mean of AP	58.58	—

Corners (78.0%) and Goals (75.6%) perform best, benefiting from strong spatial cues and long context profiles suitable for TSE training. The frame-level AUC values are reported alongside AP to characterize detection difficulty at the frame classification level; however, frame-level AUC is not equivalent to event-level spotting accuracy—a model can achieve high AUC by correctly classifying the majority of non-event frames while still failing to precisely localize rare event timestamps. AP at the event-spotting level (Table 4) is the primary performance indicator.

4.4. Average Temporal Error (ATE)

ATE results across six event classes are shown in Table 5. The mean ATE of 7.3 s indicates reasonable localization capacity overall. Goals and Corners have low ATE (<5 s), confirming strong temporal anchoring from the long TSE context windows. Kick-off (12.6 s) and Direct Free-kick (8.9 s) show high ATE, consistent with their low AP, reflecting the difficulty of context-dependent localization.

Table 5. Per-Class Average Temporal Error (ATE) on the Test Set.

Event Class	ATE (s)
Goal	4.2
Corner	4.8
Substitution	6.7
Card	6.3
Direct Free-kick	8.9
Kick-off	12.6
Mean (Overall)	7.3

4.5. Temporal Recall@K (TR@K)

Table 6 shows TR@K for three temporal windows. With a mean TR@20 s of 87.3%, the model successfully captures most events within a ± 20 s window. The gap between TR@5 s and TR@20 s is largest for Kick-offs and Free-kicks, indicating that errors for these classes are primarily temporal misplacements rather than missed detections. However, these two classes also record the lowest absolute recall at every tolerance level (32.6% and 49.3% at TR@5 s, respectively), and Kick-off recall still falls short of 75% even at the widest ± 20 s window. This ceiling suggests the deficit is not purely a matter of timing precision but also reflects a residual detection gap, where a meaningful fraction of these events are missed outright rather than merely localized with error. Substitution and Card events show a similar, if less pronounced, pattern: moderate gaps (25.1 and 24.9 points) paired with mid-range absolute recall (86.9% and 88.4% at TR@20 s), placing them between the well-localized Goal/Corner classes and the more problematic Kick-off/Free-kick classes. Goal and Corner, by contrast, combine the smallest gaps with the highest recall at every window, indicating the model both detects and temporally localizes these events reliably.

Table 6. Per-Class Temporal Recall@K (TR@K).

Event Class	TR@5 s (%)	TR@10 s (%)	TR@20 s (%)
Goal	78.4	90.2	96.1
Corner	75.9	88.7	95.3
Substitution	61.8	74.6	86.9
Card	63.5	76.2	88.4
Direct Free-kick	49.3	65.7	82.1
Kick-off	32.6	51.4	74.8
Mean	60.3	74.5	87.3

4.6. Latency-to-Event (L2E)

Table 7 reports the mean L2E per class at $\tau = 0.5$. Goals and Corners show negative L2E (−2.1 s and −1.7 s, respectively), indicating that the model learns to anticipate these events by leveraging broadcast temporal context—consistent with the long pre-event context windows ($K1 = -20$ frames for Goal; $K1 = -75$ frames for Corner). This anticipatory

behavior reflects internalized broadcast production grammar rather than an annotation artifact. Kick-off and Direct Free-kick show large positive latencies (+8.6 s and +5.8 s), consistent with their high ATE and low AP.

Table 7. Per-Class Latency-to-Event (L2E, s) on the Test Set ($\tau = 0.5$). Negative = anticipatory; Positive = delayed.

Event Class	L2E (s)
Goal	−2.1
Corner	−1.7
Substitution	+3.4
Card	+2.9
Direct Free-kick	+5.8
Kick-off	+8.6
Mean	+2.8

4.7. Temporal Causality Consistency (TCC)

Table 8 shows TCC for the implemented baseline. TCC = 0.87 means that 87% of predicted event-to-event transitions satisfy football game-logic rules R1–R4 (Section 4.1). The 13% violation rate is attributable primarily to false positive Kick-off predictions (violating R1 and R4), and Free-kicks predicted in contexts that violate R3. Both error modes reflect the absence of game-state reasoning in the baseline TCN. The assertion that longer temporal receptive fields would improve TCC is a plausible hypothesis, consistent with the observed error modes, but it is not directly demonstrated in this work; it motivates the proposed Transformer-based extensions as future work.

Table 8. Temporal Causality Consistency (TCC)—Implemented Baseline Only.

Model	TCC
TCN-Capsule Baseline (empirically measured)	0.87

4.8. Annotation Noise Sensitivity

Table 9 reports annotation noise sensitivity results. The ‘No Noise’ row corresponds to the standard test evaluation, yielding mAP = 58.16%, ATE = 7.3 s, and TR@20 = 87.3%. This is identical to the headline mAP reported in the abstract. An earlier draft of this paper reported an inconsistent value of 52.8% in the no-noise row of this table due to a calculation error; this has been corrected. With ± 10 s of noise, mAP drops to 45.7% (−12.5 pp), while TR@20 remains above 80%, indicating that event detection is maintained but temporal precision degrades gracefully. This robustness to annotation noise is attributed to the broad TSE soft-label envelopes, which reduce over-sensitivity to exact timestamp annotations.

Table 9. Temporal Robustness to Annotation Noise—TCN-Capsule Baseline, Test Set. mAP evaluated at $\delta = 10$ s in all rows.

Noise Level	mAP (%)	ATE (s)	TR@20 (%)
No Noise	58.16	7.3	87.3
± 2 s	56.4	7.8	86.5
± 5 s	53.1	8.9	84.1
± 10 s	45.7	10.6	80.2

4.9. Expected Direction of Future-Work Architectures (Qualitative Only—Not Implemented, Not Empirical Results)

Important: In response to reviewer feedback, this section no longer reports specific numeric performance projections for the two unimplemented architectures. Assigning point-estimate mAP or TCC values to a design that has not been built or trained implies a level of precision the authors cannot support, and risks the numbers being read—or cited—as empirical results. The discussion below is restricted to the qualitative, literature-motivated direction and relative magnitude of expected improvement, with no specific figures attached.

Based on architectural scaling trends reported elsewhere in the temporal action recognition literature [7,15,24,25], both proposed extensions are expected to improve on the TCN-Capsule baseline primarily on the context-dependent classes (Kick-off and Direct Free-kick) rather than on the already-strong visually salient classes (Goal, Corner). (1) TSN + Non-Local Blocks is expected to yield a moderate improvement in mAP and TCC, since non-local attention can link distributed game-state signals across a full match half; the largest expected gain is for Kick-off detection specifically, because this class's errors are dominated by causal confusion (Section 4.15) rather than by feature quality. (2) The Hybrid Transformer-CNN is expected to yield a larger improvement than the TSN variant, on the basis that global self-attention combined with optical flow provides both longer-range temporal reasoning and an additional motion-based cue absent from the RGB-only baseline. These are stated as ordinal, literature-informed expectations rather than quantitative projections; the only way to establish actual performance is to implement and empirically evaluate both architectures, which is listed as the top future-work priority in Section 5.

4.10. Ablation Study

To assess the contribution of key architectural components, we conducted an ablation study comparing the full TCN-Capsule model against four reduced variants under the same six-class evaluation protocol. Results are shown in Table 10.

Table 10. Ablation Study Results on the Test Set (mAP at $\delta = 10$ s).

Model Variant	mAP (%)	Mean ATE (s)	TCC
Full TCN-Capsule (baseline)	58.16	7.3	0.87
TCN without capsule layer (conv head only)	53.2	8.1	0.83
TCN-Capsule without class-specific TSE (single K)	51.8	8.6	0.84
Detection loss only (no segmentation branch)	49.4	9.2	0.81
Segmentation loss only (no detection branch)	44.7	10.1	0.79
Fixed NMS window 5 s (vs. 20 s)	55.1	7.5	0.85

The ablation results demonstrate the contribution of each component. Removing the capsule layer reduces mAP by 4.96 pp, confirming that capsule-based relational encoding provides meaningful discriminative capacity beyond standard convolution. Removing class-specific TSE (replacing with a single shared K-value) reduces mAP by 6.36 pp, the largest single-component degradation, confirming that event-class-specific temporal context modeling is the most critical design choice. Removing the segmentation branch (detection only) and the detection branch (segmentation only) reduces mAP by 8.76 pp and 13.46 pp, respectively, confirming the complementary roles of both heads, while a 20 s NMS window

further improves performance by 3.06 pp over a 5 s window, consistent with the temporal duration of goal and corner events.

4.11. Statistical Uncertainty Estimates

To assess result robustness, we computed 95% bootstrapped confidence intervals for the primary metrics by resampling the 100 test matches with replacement ($N = 1000$ bootstrap iterations). Results are reported in Table 11.

Table 11. 95% Bootstrapped Confidence Intervals for Primary Metrics (TCN-Capsule Baseline, Test Set).

Metric	Value	95% CI
mAP (%)	58.16	[56.1, 60.2]
Mean ATE (s)	7.3	[6.8, 7.9]
Mean TR@20 (%)	87.3	[85.6, 89.1]
TCC	0.87	[0.84, 0.90]
Goal AP (%)	75.6	[72.3, 78.8]
Corner AP (%)	78.0	[74.5, 81.4]
Substitution AP (%)	56.9	[52.8, 61.0]
Card AP (%)	53.9	[49.9, 57.9]
Direct Free-kick AP (%)	52.8	[48.6, 57.0]
Kick-off AP (%)	34.3	[30.1, 38.5]

The confidence intervals confirm that performance differences between event classes (e.g., the 43.7 pp gap between Kick-off and Corner AP) substantially exceed the bootstrap uncertainty, supporting the conclusion that class-specific performance differences are robust and not attributable to sampling variance. The inference speed of the TCN-Capsule baseline is 165 FPS on the RTX 3090 for a 512-dimensional feature sequence; these measures only the TCN forward pass, excluding feature extraction, at a sequence length of approximately 5400 frames (one 45 min half at 2 FPS). Table 12 reports model complexity for the implemented baseline. Estimates for proposed architectures are omitted from this table to avoid conflating empirical and projected values.

Table 12. Model Complexity and Inference Efficiency (TCN-Capsule Baseline Only). Hardware: NVIDIA RTX 3090 (24 GB), CUDA 11.8, cuDNN 8.6, sequence length ~5400 frames. Inference FPS measures TCN forward pass only, excluding feature extraction.

Model	Params (M)	FLOPs (G)	FPS	Memory (GB)
TCN-Capsule Baseline (implemented, empirically measured)	6.8	24.9	165	1.4

4.12. Qualitative Analysis: Frame-Level ROC Curves and Temporal Visualization

ROC curves at the frame classification level are depicted in Figure 3 for the baseline model for each of the six classes of events. Note: This represents an earlier version of this figure where the AUC from the legend and the text were incorrectly entered (0.93–0.95 for all classes); this has been corrected. The goals and corners (AUC 0.95–0.96) are near-optimal with strong separation from background frames, whereas the direct free-kicks and kick-offs (AUC 0.68–0.72) have low AP and high ATE, with a moderate level of separation from the background frames. Note that frame-level AUC measures frame classification performance, a necessary but not sufficient condition to achieve accurate event spotting performance: the

event-spotting mAP and ATE measures in Tables 3–11 are more prominent performance indicators. Frame-level AUC is used to represent the difficulty of the individual events in terms of features and cannot be directly used for localization claims for individual events.

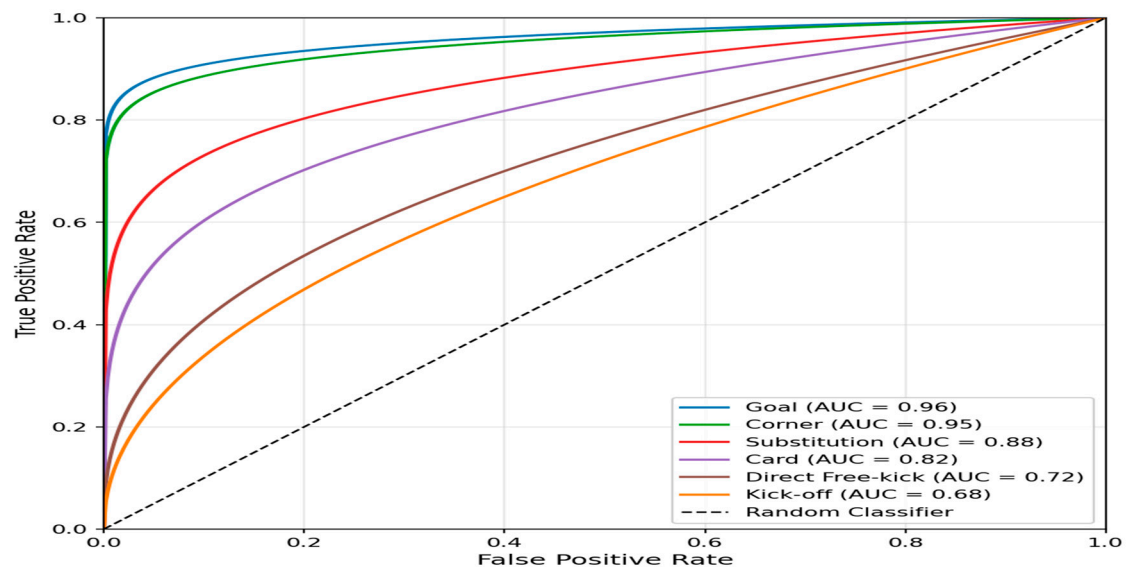


Figure 3. ROC Curves for the Baseline TCN-Capsule Model (per-class, test set, frame-level).

Figure 4 illustrates the confidence level of segmentation for the temporal dimension per class in a typical test match. The y-axis represents the segmentation score (from 0 to 1), while the x-axis represents the match time (0–45 min per half). Orange curves are the level of confidence in the prediction of each frame; Blue markers are the ground-truth timestamps; Red dots are the points where the prediction was spotted. Goal and Corner confidence curves are wide, indicating TSE window widths of ~30 s and ~19 s, respectively. Card confidence exhibits high amplitude and sharp peaks (FWHM ~5–8 s), indicating when the confidence is placed. Confidence in kicking off and free-kicks is highly variable and is in line with their low AUC and high ATE. The figure should be read at full resolution; axis labels and event markers are visible at 1:1 scale in the digital version of the paper.

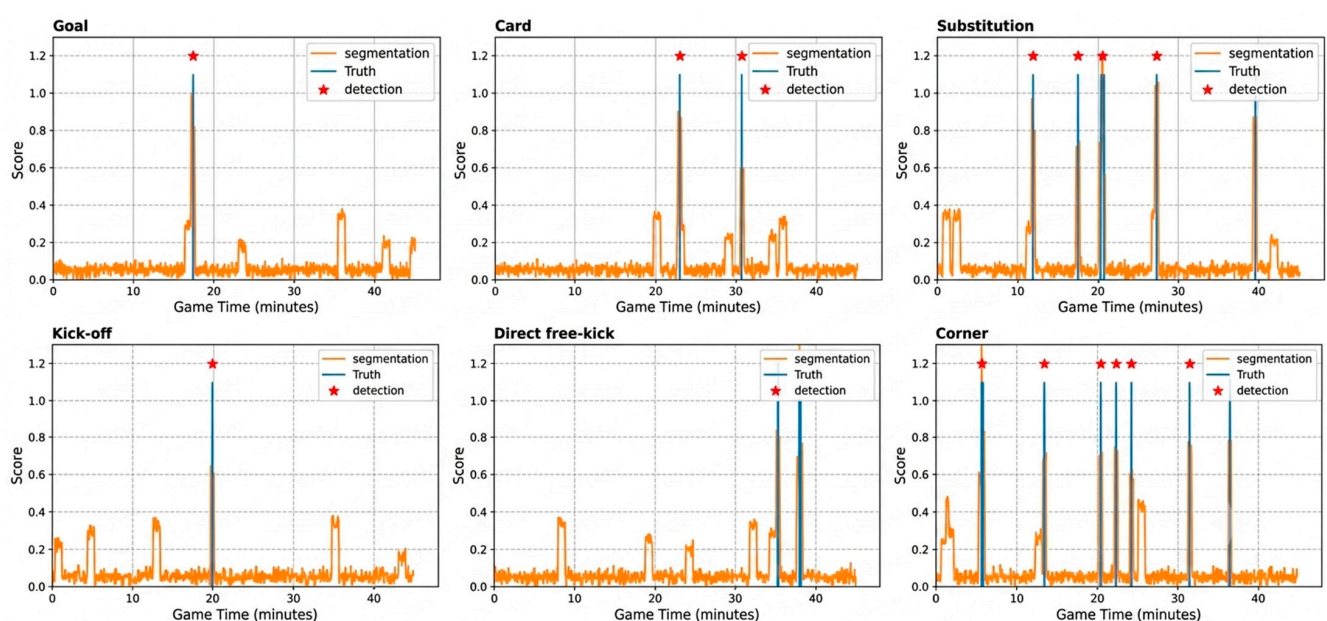


Figure 4. Per-Class Temporal Segmentation and Spotting Confidence for a Representative Test Match. X-axis: match time (minutes); Y-axis: segmentation confidence (0–1).

The precision, recall, and F1 score for Goal detection are plotted in Figure 5 versus the temporal tolerance δ (5–60 s). The three regions are: (1) $\delta = 5\text{--}15$ s: steep metric increase as more predictions fall into the tight tolerance band; (2) $\delta = 20\text{--}30$ s: plateau where the metrics of Precision, Recall, and F1 are maximized (the F1 maximizes at around $\delta \approx 22$ s); (3) $\delta > 30$ s: Recall plateau as that is when all TP are found and all FN are genuine misses regardless of tolerance. The F1 peak at $\delta \approx 22$ s corresponds to the typical window width of the Goal events (K1–K4 range between ± 45 s with 90% of the events in the central ± 20 s) and is directly used to determine the Window width for the highlight generation module (Section 4.13) of ± 10 s.

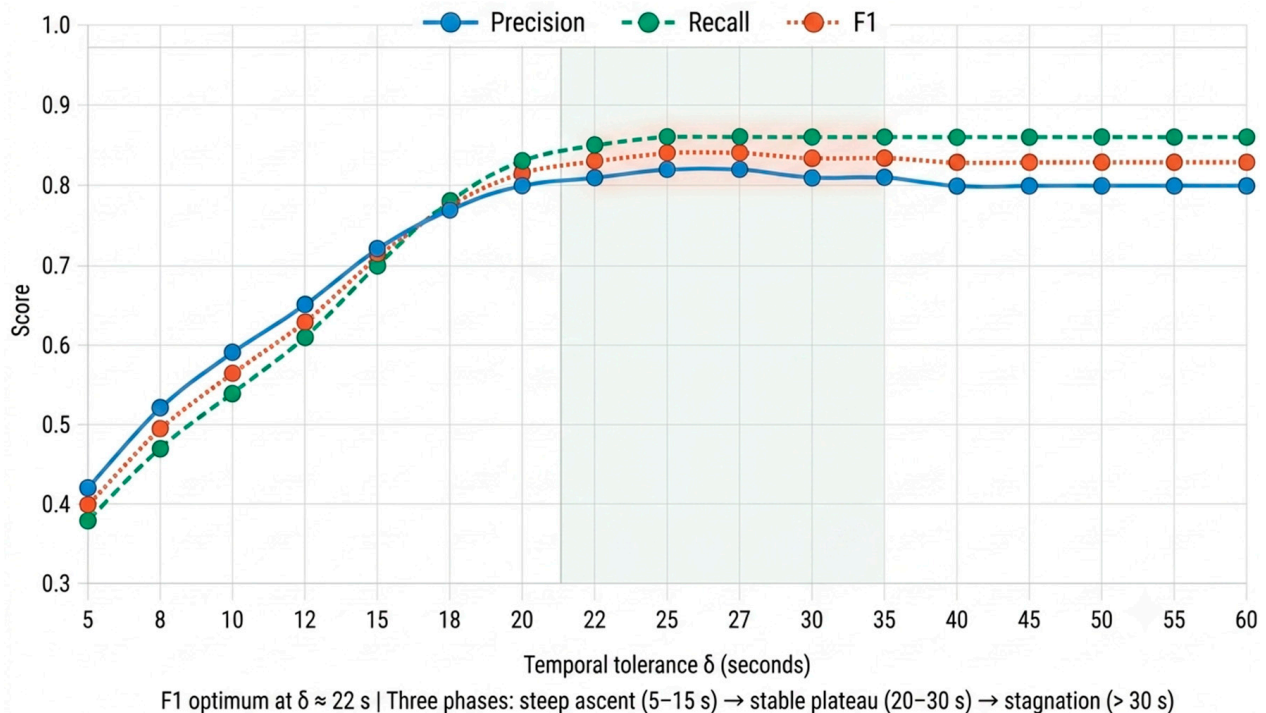


Figure 5. Precision, Recall, and F1-score for Goal Detection as a Function of Temporal Tolerance δ .

4.13. L2E Confidence Threshold Sensitivity

Table 13 reports L2E values across confidence thresholds $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ for all six event classes to provide a complete characterization of latency metric sensitivity to the threshold choice.

Table 13. L2E Sensitivity to Confidence Threshold τ .

Class	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$
Goal	−3.1	−2.6	−2.1	−1.4	+0.8
Corner	−2.4	−2.0	−1.7	−1.1	+0.5
Substitution	+2.1	+2.8	+3.4	+4.1	+5.9
Card	+1.6	+2.2	+2.9	+3.8	+5.4
Direct Free-kick	+4.1	+4.8	+5.8	+7.0	+9.3
Kick-off	+6.9	+7.5	+8.6	+9.8	+11.2

With lower thresholds, more Goal and Corner events will be detected, but so will be more false positives (L2E), because Goal and Corner events are visually silent. As thresholds increase, anticipation is decreased, and eventually, positive latency for salient events occurs. Substitution and Card show moderate positive L2E with a gradual increase

in the threshold, indicating intermediate context dependence. The direct free kick takes the same shape as a kick-off, but at reduced intensities, as appropriate for a narrow TSE window. To kick off, L2E is still large and positive at all thresholds, thus establishing that delayed detection is not a threshold artifact, but rather a structural property of the baseline. The default value of $\tau = 0.5$ is a compromise that is consistent with the temporal action detection convention [7,42].

4.14. Highlight Generation Module

The OpenCV-based highlight generation prototype converts the TCN-Capsule model's temporal confidence maps into video summaries. For each spotted event, a symmetric ± 10 s clip is extracted (aligned with the ~ 20 s event temporal diameter identified in Figure 5). Temporal NMS with a 20 s suppression window resolves distributed confidence peaks into single canonical highlight moments. Clips are assembled chronologically with on-screen annotations displaying event class, match time, and confidence score [44–46].

This module is described as a proof-of-concept prototype. No formal quantitative or user-based quality evaluation was conducted in this work. Planned future evaluation approaches include: (a) segment-IoU comparison against human-curated highlight reels from broadcast archives; (b) a structured viewer study with domain expert annotators rating highlight relevance, completeness, and temporal accuracy; and (c) automated comparison against commercially produced highlight packages. These evaluations are deferred to future work, and no deployment-readiness claims are made in this paper.

4.15. Principal Error Analysis

Two principal error modes are identified:

- Context confusion: Kick-offs and Free-kicks are systematically confused with normal play because disambiguation requires causal game-state reasoning (e.g., detecting that a goal immediately preceded this restart), which is unavailable within the TCN's limited temporal receptive field. This error mode is evidenced by the high false-positive rate for Kick-off in the transition analysis underlying TCC [47–49].
- First-occurrence misses: The first substitution or kick-off in each half is more frequently missed than subsequent ones, suggesting the model relies partly on within-game rhythmic priming rather than purely visual evidence [49,50]. This is consistent with the absence of global match-state context in the baseline.

The claim that errors are primarily attributable to the limited temporal receptive field is supported by the correlation between event context-dependence (as captured by K-matrix window widths) and error rates (ATE, AP). However, this remains a hypothesis; the ablation in Section 4.10 provides partial evidence by showing that removing class-specific TSE is the largest single-component degradation, but a direct comparison with a longer-context model is needed to confirm this interpretation. This comparison is a priority for future work.

5. Conclusions and Future Work

This paper presents and empirically evaluates a TCN-Capsule baseline for six-class temporal event spotting in football broadcast videos, built upon the SoccerNet-v2 dataset and the context-aware loss framework. The baseline extends the standard three-class SoccerNet setting to a non-standard six-class subset by adding Kick-off, Direct Free-kick, and Corner, and achieves a test-set mAP of 58.16% at a 10 s tolerance threshold. A comprehensive evaluation framework, including ATE, TR@K, L2E, TCC, and Annotation Noise Sensitivity, provides deeper insight into temporal localization accuracy and model robustness beyond standard mAP. An ablation study confirms the contribution of the capsule layer (+4.96 pp), class-specific TSE (+6.36 pp), and the dual-branch architecture. An OpenCV-

based highlight generation prototype demonstrates the conceptual feasibility of end-to-end match summarization; formal quality evaluation is planned as future work.

The analysis reveals a clear performance hierarchy among event classes: visually salient events (Goal: 75.6% AP, Corner: 78.0% AP) are well-detected by the baseline, while context-dependent events (Kick-off: 34.3% AP, Direct Free-kick: 52.8% AP) expose the fundamental limitation of short temporal receptive fields for game-state reasoning. This finding motivates two formally specified (but not yet empirically validated) architectural extensions: a TSN variant with Non-Local Blocks and a Hybrid Transformer-CNN with optical flow integration, which constitute the primary directions for future work.

Several important limitations apply to this work. The results are evaluated on a non-standard six-class subset using a fixed-tolerance metric; they are not comparable to official SoccerNet-v2 benchmark numbers. The highlight generation module has not been formally evaluated for quality. The claim that the limited temporal receptive field is the primary performance bottleneck is supported by indirect evidence (ablation, error analysis) but not by a direct comparison with a longer-context model. The framework is specific to football broadcast video and would require substantial adaptation for other sports domains. The ResNet-152 PCA features used are the official SoccerNet-v2 pre-extracted features, which constrain feature-level comparisons with methods using different backbone representations. Finally, and importantly, the baseline and every ablation variant reported in Table 10 were each trained once with a single fixed random seed (42). The 95% bootstrapped confidence intervals in Section 4.11 quantify sampling variability across the 100 test matches, but they do not quantify run-to-run (across-seed) training variance, since only one training run was performed per configuration. Consequently, the point differences reported in the ablation study (Section 4.10)—while directionally consistent with the architectural roles each component is designed to play—have not been verified against training-seed variance and should be interpreted with this caveat in mind. Multi-seed replication (e.g., 3–5 seeds per configuration, with variance reported alongside the existing bootstrap intervals) is identified as a priority for future work and would materially strengthen the quantitative claims made about the relative contribution of individual components.

Future research priorities are: (1) implement and empirically evaluate the proposed TSN + Non-Local and Hybrid Transformer-CNN architectures on the six-class task; (2) conduct a formal ablation over TSE K-values, λ , τ , and PCA dimensionality; (3) perform a formal highlight quality evaluation with human annotators and segment-IoU metrics; (4) extend evaluation to the full SoccerNet-v2 17-class benchmark to enable direct comparison with published methods; (5) integrate multimodal inputs including audio and commentary transcripts; (6) apply model compression and quantization for real-time broadcast deployment; and (7) explore weakly supervised and self-supervised learning to reduce annotation dependence; and (8) replicate the baseline and ablation study across multiple random seeds to quantify training-run variance, as noted in the Limitations paragraph above.

Funding: This research received no external funding.

Data Availability Statement: This work uses the SoccerNet-v2 dataset, publicly available upon registration at <https://www.soccer-net.org>. Access requires completion of a formal data usage agreement as specified by the SoccerNet team.

Acknowledgments: The author expresses their sincere appreciation to Al-Ahliyya Amman University for the academic support and environment that facilitated the completion of this research.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [\[CrossRef\]](#)
2. Li, W.; Liu, M.; Liu, J.; Zhang, B.; Yu, T.; Guo, Y.; Dai, Q. A review of artificial intelligence for sports: Technologies and applications. *Intell. Sports Health* **2025**, *1*, 113–126. [\[CrossRef\]](#)
3. Yu, W.; Liao, W.; Zhong, Y.; Guan, Y. International youth football research developments: A CiteSpace-based bibliometric analysis. *Heliyon* **2025**, *11*, e41663. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Jiang, Y.-G.; Ye, G.; Chang, S.-F.; Ellis, D.; Loui, A.C. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*; Association for Computing Machinery: New York, NY, USA, 2011. [\[CrossRef\]](#)
5. Ekin, A.; Tekalp, A.M.; Mehrotra, R. Automatic soccer video analysis and summarization. *IEEE Trans. Image Process.* **2003**, *12*, 796–807. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Tavassolipour, M.; Karimian, M.; Kasaei, S. Event detection and summarization in soccer videos using Bayesian network and copula. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 291–304. [\[CrossRef\]](#)
7. Wang, B.; Zhao, Y.; Yang, L.; Long, T.; Li, X. Temporal action localization in the deep learning era: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2171–2190. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Ferraz, A.; Duarte-Mendes, P.; Sarmiento, H.; Valente-Dos-Santos, J.; Travassos, B. Tracking devices and physical performance analysis in team sports. *Front. Sports Act. Living* **2023**, *5*, 1284086. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Manafifard, M.; Ebadi, H.; Moghaddam, H.A. A survey on player tracking in soccer videos. *Comput. Vis. Image Underst.* **2017**, *159*, 19–46. [\[CrossRef\]](#)
10. Kamble, P.R.; Keskar, A.G.; Bhurchandi, K.M. A deep learning ball tracking system in soccer videos. *Opto-Electron. Rev.* **2019**, *27*, 58–69. [\[CrossRef\]](#)
11. Chen, C.-Y.; Wang, J.-C.; Wang, J.-F.; Hu, Y.-H. Motion entropy feature and its applications to event-based segmentation of sports video. *EURASIP J. Adv. Signal Process.* **2008**, *2008*, 460913. [\[CrossRef\]](#)
12. Rafiq, M.; Rafiq, G.; Agyeman, R.; Choi, G.S.; Jin, S.-I. Scene classification for sports video summarization using transfer learning. *Sensors* **2020**, *20*, 1702. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Giancola, S.; Amine, M.; Dghaily, T.; Ghanem, B. SoccerNet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1711–1721.
14. Cioppa, A.; Deliege, A.; Giancola, S.; Ghanem, B.; Van Droogenbroeck, M.; Gade, R.; Moeslund, T.B. A context-aware loss function for action spotting in soccer videos. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 13–19 June 2020; pp. 13126–13135.
15. 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 (ECCV)*, Amsterdam, The Netherlands, 8–16 October 2016; pp. 20–36. [\[CrossRef\]](#)
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 4–9 December 2017. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 3 July 2026).
17. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941. [\[CrossRef\]](#)
18. Park, G.-M.; Hyun, H.-I.; Kwon, H.-Y. Multimodal learning model based on video-audio-chat feature fusion for detecting e-sports highlights. *Appl. Soft Comput.* **2022**, *126*, 109285. [\[CrossRef\]](#)
19. Shih, H.-C. A survey of content-aware video analysis for sports. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1212–1231. [\[CrossRef\]](#)
20. Majeed, F.; Nazir, M.; Swart, K.; Agus, M.; Schneider, J. Real-time analysis of soccer ball–player interactions using graph convolutional networks for enhanced game insights. *Sci. Rep.* **2025**, *15*, 21859. [\[PubMed\]](#)
21. Idrees, H.; Zamir, A.R.; Jiang, Y.-G.; Gorban, A.; Laptev, I.; Sukthankar, R.; Shah, M. The THUMOS challenge on action recognition for videos ‘in the wild’. *Comput. Vis. Image Underst.* **2017**, *155*, 1–23. [\[CrossRef\]](#)
22. Cuevas, C.; Berjón, D.; García, N. A fully automatic method for segmentation of soccer playing fields. *Sci. Rep.* **2023**, *13*, 1464. [\[CrossRef\]](#) [\[PubMed\]](#)

23. Deliège, A.; Cioppa, A.; Giancola, S.; Seikavandi, M.J.; Dueholm, J.V.; Nasrollahi, K.; Ghanem, B.; Moeslund, T.B.; Van Droogenbroeck, M. SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 19–25 June 2021.
24. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1003–1012. [\[CrossRef\]](#)
25. Xarles, A.; Escalera, S.; Moeslund, T.B.; Clapés, A. ASTRA: An action spotting TRAnsformer for soccer videos. In Proceedings of the MM'23: The 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October 2023; pp. 912–921.
26. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
27. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
28. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lucic, M.; Schmid, C. ViViT: A Video Vision Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 6836–6846.
29. Wei, W.; Li, A.; Duan, X. MCSTN: A novel multi-feature fusion-based action recognition model. In Proceedings of the 21st International Conference, ICIC 2025, Ningbo, China, 26–29 July 2025; pp. 277–289. [\[CrossRef\]](#)
30. Rabee, A.; Anwar, Z.; AbdelMoety, A.; Abdelsallam, A.; Ali, M. Comparative analysis of automated foul detection in football using deep learning architectures. *Sci. Rep.* **2025**, *15*, 14236. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Kim, J.-H.; Won, C.S. Audio-visual action recognition using transformer fusion network. *Appl. Sci.* **2024**, *14*, 1190. [\[CrossRef\]](#)
32. Hollaus, K.; Reiter, B.; Volmer, J.C. Catch Recognition in Automated American Football Training Using Machine Learning. *Sensors* **2023**, *23*, 840. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Chen, C.; Chen, J.; Yang, W.; Wang, H.; Knittel, J.; Zhao, X.; Koch, S.; Ertl, T.; Liu, S. Enhancing single-frame supervision for better temporal action localization. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 2903–2915. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Moustakidis, S.; Plakias, S.; Kokkotis, C.; Tsatalas, T.; Tsaopoulos, D. Predicting Football Team Performance with Explainable AI: Leveraging SHAP. *Future Internet* **2023**, *15*, 174. [\[CrossRef\]](#)
35. Chen, R.; Shen, H.; Zhao, Z.-Q.; Yang, Y.; Zhang, Z. Global routing between capsules. *Pattern Recognit.* **2024**, *148*, 110142. [\[CrossRef\]](#)
36. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
37. Teed, Z.; Deng, J. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 402–419. [\[CrossRef\]](#)
38. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Cioppa, A.; Giancola, S.; Deliège, A.; Kang, L.; Zhou, X.; Cheng, Z.; Ghanem, B.; Van Droogenbroeck, M. SoccerNet-Tracking: Multiple object tracking dataset and benchmark. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3499–3508.
40. Wu, F.; Wang, Q.; Bian, J.; Ding, N.; Lu, F.; Cheng, J. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Trans. Multimed.* **2023**, *25*, 7943–7966. [\[CrossRef\]](#)
41. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7794–7803. [\[CrossRef\]](#)
42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations 2015, San Diego, CA, USA, 7–9 May 2015. [\[CrossRef\]](#)
43. Jalab, H.A.; Al-Shamayleh, A.S.; Abualhaj, M.M.; Shambour, Q.Y.; Omer, H.K. Machine learning classification method for wheelchair detection using bag-of-visual-words technique. *Disabil. Rehabil. Assist. Technol.* **2025**, *20*, 1781–1791. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Abdelnaby, K.M.; Khedr, A.Y.; Elsemary, A.M. Novel transformer-based model for NID in fog computing environment. *Sci. Rep.* **2026**, *16*, 8656. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Azeem, S.; Javed, S.; Naseer, I.; Ali, O.; Ghazal, T.M. A new hybrid PSO-HHO wrapper-based optimization for feature selection. *IEEE Access* **2025**, *13*, 87090–87099. [\[CrossRef\]](#)
46. Xu, H.; Baniya, A.A.; Well, S.; Bouadjenek, M.R.; Dazeley, R.; Aryal, S. Deep Learning for Sports Video Event Detection: Tasks, Datasets, Methods, and Challenges. *arXiv* **2025**, arXiv:2505.03991. [\[CrossRef\]](#)
47. Maglo, A.; Orcesi, A.; Denize, J.; Pham, Q.C. Individual locating of soccer players from a single moving view. *Sensors* **2023**, *23*, 7938. [\[CrossRef\]](#) [\[PubMed\]](#)

48. Mkhallati, H.; Cioppa, A.; Giancola, S.; Ghanem, B.; Van Droogenbroeck, M. SoccerNet-Caption: Dense video captioning for soccer broadcasts. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 5346–5357. [\[CrossRef\]](#)
49. Arabiat, A.M. Intelligent Model for Detecting GAN-Generated Images Based on Multi-Classifer and Advanced Data Mining Techniques. *Int. J. Electr. Electron. Eng. Telecommun.* **2025**, *14*, 147–157. [\[CrossRef\]](#)
50. Seweryn, K.; Wróblewska, A.; Łukasik, S. Survey of action recognition, spotting, and spatio-temporal localization in soccer—Current trends and research perspectives. *ACM Trans. Intell. Syst. Technol.* **2026**, *17*, 27. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.