

PoseConvSpot: Overcoming Domain Shifts in Precise Event Spotting for Sports Videos via Pose-Only Spatio-Temporal Regression

Noâm Boussouf
Keio University
Yokohama, Japan
noam.boussouf@keio.jp

Adrien Maglo
Université Paris-Saclay, CEA, List
F-91120, Palaiseau, France
adrien.maglo@cea.fr

Mohamed Chaouch
Université Paris-Saclay, CEA, List
F-91120, Palaiseau, France
mohamed.chaouch@cea.fr

Paul Pham Dang
Keio University
Yokohama, Japan
paul.phamdang@keio.jp

Hideo Saito
Keio University
Yokohama, Japan
hs@keio.jp

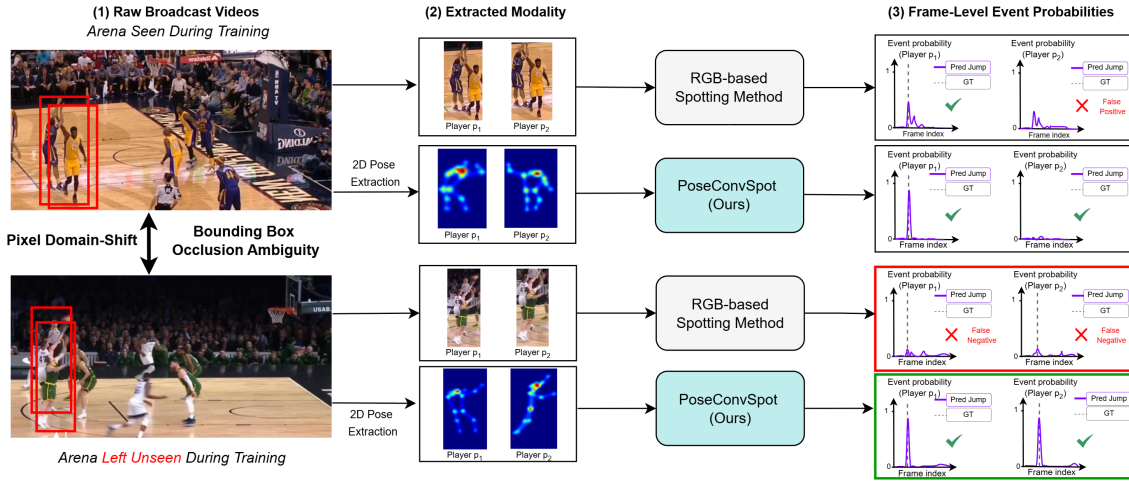


Figure 1: Robust precise event spotting under domain shifts and occlusions. Top: RGB baselines overfit to training arenas, failing on unseen courts and during player overlaps. Bottom: By projecting skeletons into video heatmaps, PoseConvSpot isolates athletes, yielding stable predictions and a +14.1% mAP gain in novel environments over state-of-the-art Spotting Methods [23].

Abstract

Precise Event Spotting (PES) demands frame-level accuracy for localized actions in sports analytics. Current methodologies overwhelmingly rely on end-to-end pipelines trained directly on raw visual inputs. While effective in controlled settings, these models fail to scale in dense multi-player scenarios and are severely vulnerable to pixel-level domain shifts, demanding massive datasets to overcome variations in court layouts and camera angles. To address these limitations, we propose *PoseConvSpot*, a pose-only spatio-temporal regression framework tailored for high-fidelity event spotting. Operating exclusively on 2D joint coordinates, our pipeline projects skeletons into spatio-temporal heatmaps that natively isolate individual actors. By combining directional *Strip Pooling* with a *Shared-Head GRU* and a continuous *Soft Gamma Loss*,

PoseConvSpot accurately spots biomechanical events without relying on fine-grained visual cues traditionally deemed essential for precise spotting. Evaluations on a multi-sport jump benchmark show that our framework achieves performance competitive with state-of-the-art RGB models while delivering remarkably superior cross-arena robustness (+14.1% absolute mAP@3 gain on unseen courts) and stable, peak-focused predictions without heuristic post-processing. Project Page: noambsf.github.io/PoseConvSpot.

CCS Concepts

• Computing methodologies → Activity recognition and understanding.

ACM Reference Format:

Noâm Boussouf, Adrien Maglo, Mohamed Chaouch, Paul Pham Dang, and Hideo Saito. 2026. PoseConvSpot: Overcoming Domain Shifts in Precise Event Spotting for Sports Videos via Pose-Only Spatio-Temporal Regression. In *9th Int. Workshop on Multimedia Content Analysis in Sports (MMSports '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3841455.3841529>



This work is licensed under a Creative Commons Attribution 4.0 International License. *MMSports '26*, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2944-7/26/11
<https://doi.org/10.1145/3841455.3841529>

1 Introduction

Precise Event Spotting (PES) is the task of localizing instantaneous events with frame-level accuracy. In the context of dynamic team sports, identifying explosive, transient events is of paramount importance. Achieving such high temporal precision for jump spotting is crucial, as it serves as the mandatory foundation for complex downstream applications, most notably Jump-Aware Game-State Reconstruction [13] from a single moving camera as well as 3D Reconstruction of Sports Sequences from monocular videos. However, isolating these micro-events within a continuous video stream presents specific difficulties. While current video understanding architectures effectively recognize continuous, long-form actions [12, 19], they struggle with high-fidelity temporal localization. Unlike traditional Temporal Action Localization (TAL) [25] or general Event Spotting (ES) [26], which often tolerate coarse temporal boundaries, PES strictly demands extremely low error at the exact frame level.

Current state-of-the-art methodologies in precise event spotting predominantly rely on continuous RGB video streams processed through end-to-end network pipelines. These frameworks typically formulate the task as frame-level classification, categorizing each frame as either a target event or background by leveraging long-range temporal dependencies alongside multi-scale and/or multi-resolution feature representations [7, 15, 23, 24]. However, operating directly on raw pixel modalities introduces severe vulnerabilities to domain shifts, resulting in degraded robustness when models are deployed across unconstrained, unknown environments. Variations in viewpoint, lighting, and arena backgrounds require extensive data collection and annotation efforts. Furthermore, these architectures are fundamentally limited to global event detection. When multiple localized micro-events occur concurrently, such as different players jumping simultaneously, distinguishing individual actions requires independently tracking and processing separate player tracklets. This reliance on external tracking pipelines makes the system highly sensitive to player occlusions and spatial overlaps while further hindering scalability.

To address these limitations, we introduce **PoseConvSpot**, a pose-only framework that enables simultaneous, localized precise event spotting. By operating solely on readily available human pose priors, our method leverages the strong generalization capabilities of modern pose estimators to natively isolate interacting athletes and becomes inherently immune to pixel-level domain shifts, maintaining highly confident predictions even in completely unseen arenas (Figure 1). As illustrated in our overall architecture (Figure 2), rather than processing raw images, PoseConvSpot takes as input the raw 2D joint coordinates of each player. To eliminate absolute spatial displacement and camera ego-motion, coordinates are normalized via a horizontal pelvis-centered alignment and projected into a continuous sequence of spatio-temporal skeleton heatmaps. These stacked 3D volumes are subsequently processed by a 3D convolutional backbone coupled with a head architecture tailored to capture the specific spatial movements and temporal dynamics of jumping events.

Specifically, we resolve the spatial and temporal challenges of PES through two key architectural innovations. First, to effectively capture the distinct kinematic signatures of a jump, the network

must identify coordinated body parts across the entire spatial layout. To achieve this, we integrate a *Strip Pooling* module. Unlike standard square pooling windows that restrict receptive fields to local regions, Strip Pooling aggregates spatial features along complete and independent vertical and horizontal axes. This formulation allows the network to effectively capture long-range axial dependencies, most notably, the full vertical alignment and structural extension of the athlete’s body during the ascending phase of a jump, yielding a highly discriminative spatial representation. Second, while the vast majority of current PES methods formulate the task as a discrete frame-classification problem, which suffers from severe class imbalance, we model it as a continuous temporal regression task. To effectively optimize this formulation, we deploy a unified *GRU-based Temporal Head* trained with a novel application of a *Soft Gamma Loss*. This joint architecture forces the network to regress continuous probability distributions for target events. Consequently, this design not only mitigates sparsity issues but significantly boosts the model’s reliability for practical deployment, yielding high performance on strict, operational metrics (F_1 -score) without requiring hyperparameter-sensitive post-processing steps like Non-Maximum Suppression, which are strictly required in classification frameworks to prevent performance collapse.

Through extensive experiments, we demonstrate that despite the complete absence of fine-grained visual (RGB) textures, our pose-only approach delivers performance competitive with or superior to heavy state-of-the-art architectures for jump spotting. More importantly, our framework exhibits exceptional data efficiency and cross-arena generalization capabilities, maintaining high temporal fidelity even in severely data-constrained regimes.

Our key contributions can be summarized as follows:

- We propose a precise event spotting pipeline relying exclusively on low-dimensional kinematic representations. By operating strictly on 2D poses, our framework inherently isolates individual actors to resolve dense multi-player interactions and demonstrates remarkable robustness to pixel-level domain shifts.
- We introduce a specialized spatio-temporal regression architecture tailored for PES. By combining *Strip Pooling*, which effectively captures long-range axial dependencies to identify full structural body extensions, with a *Shared-Head GRU* and a continuous *Soft Gamma Loss*, our model generates highly localized peak predictions, eliminating the need for heuristic post-processing steps such as Non-Maximum Suppression (NMS).
- We provide a rigorous empirical evaluation demonstrating that our framework achieves highly competitive results against state-of-the-art RGB-based PES models. Furthermore, we highlight the model’s superior data efficiency and generalization capabilities through *Leave-One-Arena-Out* (LOAO) cross-venue validation, proving its stability across diverse environments without requiring domain-specific adaptation.

2 Related Work

From Temporal Action Localization to Precise Event Spotting. Temporal Action Localization focuses on identifying the start and end times of actions within untrimmed videos [25]. While modern TAL architectures, often based on Transformers or temporal convolutions [3, 18, 22, 29], achieve strong performance on coarse action intervals, they are generally inadequate for capturing instantaneous precise occurrences. To address this, Precise Event Spotting emerged as a distinct paradigm targeting highly localized actions requiring frame-level accuracy, heavily popularized by datasets like SoccerNet [2, 6] and FineGym [16].

Recent state-of-the-art PES methodologies [7, 15, 20–24] overwhelmingly favor end-to-end architectures operating directly on RGB video streams. Pioneering works like E2E-Spot [7] and T-DEED [23] demonstrated that jointly learning spatial features and temporal dynamics outperform multi-stage pipelines. Subsequent advances have sought to refine this paradigm: Santra et al. [15] explicitly tackle long-range temporal dependencies and severe class imbalance, while recent redundancy-aware models like AdaSpot [24] attempt to mitigate the massive computational overhead of RGB streams by adaptively focusing on salient spatial regions.

However, this reliance on the pixel domain still introduces critical bottlenecks. First, RGB models remain highly susceptible to domain shifts (e.g., varying lighting, camera angles, and arena backgrounds for indoor sports videos), demanding massive, diverse datasets to generalize. Second, because they process the global scene, these models are not suited for independently resolving simultaneous micro-events (e.g., multiple players jumping concurrently). Finally, most RGB spotting pipelines formulate the task as frame-level classification, necessitating sensitive post-processing techniques like Non-Maximum Suppression to filter redundant predictions. Our work diverges from this paradigm by abandoning the RGB modality entirely for PES, remarkably improving results in unseen environments while natively isolating individual players.

Skeleton-Based Action Recognition. To mitigate the computational cost and domain sensitivity of RGB models, human pose sequences have been widely adopted for Action recognition [12, 19] (i.e., classifying a pre-trimmed video segment into a single category), offering a computationally efficient and environmentally robust alternative. Traditional approaches frequently employ Spatial-Temporal Graph Convolutional Networks (ST-GCN) [28] to model the inherent topological structure of the human body. Alternatively, pseudo-image representations, such as PoseConv3D [4], map the temporal and spatial dimensions into an image-like format by stacking 2D joint heatmaps across time for processing via 3D CNNs. This pseudo-image paradigm has proven highly effective, as it smoothly models spatial uncertainty and seamlessly integrates into standard convolutional pipelines without the complex message-passing bottlenecks typical of GCNs.

Despite their success, existing pose-based architectures are predominantly designed for clip-level *Action Recognition* [12]. Applying these representations to continuous, untrimmed videos for fine-grained *Precise Event Spotting* remains largely unexplored. This is likely due to the prevailing assumption that achieving frame-level accuracy strictly requires fine-grained visual cues from the RGB

environment (e.g., visual player-ground contact, precise object interactions). PoseConvSpot fundamentally challenges this intuition. By adapting a robust heatmap representation for continuous temporal regression, our pipeline demonstrates that extracting geometric deformations can be sufficient to achieve competitive frame-level precision. This paradigm shift allows our framework to deliver high temporal fidelity while remaining inherently immune to pixel-level domain shifts and naturally bypassing the spatial overlap and occlusion issues that severely hinder RGB models when processing simultaneous multi-actor events.

Jump Precise Event Spotting. In sports computer vision, explosive vertical jumps serve as an ideal benchmark for evaluating fine-grained micro-events. In dynamic team sports such as volleyball and basketball, these actions frequently occur concurrently among multiple interacting players. For RGB-based PES methods to handle such simultaneous micro-events, they are forced to process individual player tracklets independently. This heavy reliance on extracting and feeding multiple cropped RGB sequences which can cause the system to fail during player occlusions and bounding box overlaps. While early biomechanical literature primarily focused on downstream jump-centric tasks using manually pre-trimmed video clips [10, 14, 17], deploying these analytics on continuous, untrimmed broadcasts strictly requires automated spotting as a prerequisite. Recognizing this, recent frameworks for 2D Game State Reconstruction [13] and 3D Jump-Aware collective game reconstruction from a single-view camera require automated spotting pipelines. Consequently, the success of these complex downstream applications is now fundamentally bottlenecked by the accuracy, scalability, and domain-robustness of the underlying spotting engine, underscoring the need for highly efficient frameworks like PoseConvSpot.

Despite this essential need, the fine-grained temporal localization of jumps remains underexplored. Recently, the introduction of the *SportsJumpMotion* [13] dataset provided a large-scale benchmark featuring frame-accurate jump annotations for basketball and volleyball. Alongside this dataset, the authors framed jump event spotting as a sliding-window regression problem. Their method maps cropped RGB tracklets to Gaussian-smoothed probability distributions, optimized via a Mean Squared Error loss. While this avoids rigid binary classification, their approach relies on a computationally heavy Vision Transformer backbone to process short frame sequences (16) of high-resolution (224×224) RGB tracklets [13]. This necessitates dense pixel-level processing for every player, which fails to scale in multi-agent environments and inherits the same domain-sensitivity limitations as other RGB-based frameworks. Furthermore, employing a standard MSE loss for sparse event spotting is notoriously unstable, as it remains highly sensitive to severe class imbalance and hyperparameter tuning.

PoseConvSpot overcomes these limitations through a unified spatio-temporal regression design tailored for video skeletal heatmaps. By combining pelvis-aligned sequences with directional *Strip Pooling* [8], we effectively capture long-range axial dependencies to explicitly model the athlete’s full structural extension. Importantly, we pair this highly discriminative spatial representation with a Shared-Head GRU trained via a continuous *Soft Gamma Loss*. This formulation completely bypasses the need for heuristic post-processing

steps like NMS, while demonstrating that pure kinematic deformations can effectively isolate fine-grained biomechanical events without relying on environmental visual cues.

3 Proposed Method

3.1 Problem Definition

Our proposed framework, PoseConvSpot, is designed for robust, multi-actor PES. By relying exclusively on 2D geometric priors rather than dense RGB streams, the model natively isolates interacting athletes and captures the biomechanical deformations crucial for precise temporal localization. As illustrated in Figure 2, PoseConvSpot processes the data through a decoupled, streamlined pipeline.

First, if pose joints are not readily available, videos are processed through an off-the-shelf 2D pose extractor to obtain full video sequences (of length V) containing the raw 2D joint coordinates for all identified players. For each individual player tracklet, these long sequences are partitioned into fixed-length clips of T frames, which are then transformed into video pseudo-image heatmaps. This generation step applies a pelvis-centric alignment to horizontally center the action (Section 3.2), decoupling the athlete’s actual movements from global camera shifts. The resulting stacked heatmap volumes are then fed into a 3D convolutional backbone (here, ResNet3D) to extract high-dimensional spatio-temporal features. Finally, these features are processed by a unified Shared Regression Head. Within this head, a directional Strip Pooling module explicitly models vertical body elongations and horizontal structures. A subsequent GRU layer models the temporal dynamics across the sequence, and a Multi-Layer Perceptron outputs a $|\mathcal{E}|$ -dimensional vector representing the continuous, frame-level probability scores for all $|\mathcal{E}|$ target event classes (Section 3.3). This entire spatio-temporal architecture is trained end-to-end using a continuous Soft Gamma Loss (Section 3.4), further supported by specialized data augmentations directly applied to the skeletal representations.

3.2 Heatmap Generation and Augmentation

The extracted 2D human pose sequence of length V is partitioned into fixed-length clips of T frames. Raw 2D joint coordinates are inherently entangled with global camera motion, such as panning and tilting. To explicitly isolate the athlete’s biomechanical movements, we apply a 1D *Pelvis-Centric Horizontal Alignment*. For each frame, we compute the horizontal center of the pelvis (the average horizontal coordinate of the left and right hips) and translate the entire skeleton along the horizontal axis to center it.

Similar to PoseConv3D [4], the normalized discrete coordinates are then projected into pseudo-image heatmaps of spatial size (H_{htmp}, W_{htmp}) by rendering 2D Gaussian peaks at each joint location. Spatial features are then disentangled by assigning a dedicated channel to each anatomical part, yielding C channels. For a skeleton defined by J keypoints, this yields J independent channels. We also render limb heatmaps by interpolating Gaussians along the segments connecting adjacent joints, providing J additional channels. If a player temporarily disappears from the camera view, their missing joint coordinates are naturally rendered as empty channels for the affected frames.

To prevent overfitting, we apply specialized data augmentations. At training time, we perturb the coordinates with *Random Key-point Noise*. To improve robustness against player overlaps and discourage over-reliance on specific joints, we employ *Random Joint Dropping*, which artificially sets the confidence scores of randomly selected joints to zero. Finally, we incorporate *Random Scale Padding*, *Horizontal Flip* and *Random Temporal Cropping* to simulate varying camera zoom levels and shifts in the micro-event’s temporal position within the T -frame window, respectively. Data Augmentation ablation are available in Supplementary Material C.

3.3 Spatio-Temporal Regression Architecture

The heatmap volumes are first processed by a 3D convolutional backbone, specifically, we utilize the *SlowOnly* architecture [4, 5], which is based on a 3D ResNet backbone, to extract high-dimensional spatio-temporal features $\mathbf{F} \in \mathbb{R}^{T \times d_1 \times H_{out} \times W_{out}}$. To map these latent representations into precise frame-level event probabilities, we introduce a specialized spatio-temporal regression head.

Conventional spatial reduction techniques often rely on Global Average Pooling (GAP). However, GAP dilutes highly directional biomechanical signals, such as the rapid vertical extension of an athlete’s body during a jump. To overcome this limitation, we integrate a directional *Strip Pooling* module [8]. Operating independently on the spatial dimensions of each frame, Strip Pooling employs long, narrow kernels to aggregate context along pure horizontal and vertical axes. This mechanism explicitly isolates vertical elongations and horizontal structural cues independently, outputting a spatially flattened sequence of feature vectors $\mathbf{F}_{seq} \in \mathbb{R}^{T \times d_1}$.

To model the continuous kinematics of the movement, the spatial sequence is first processed through a shallow 1D temporal convolution module. Composed of two consecutive 1D convolutional layers, each followed by Batch Normalization [9] and ReLU activation, this module acts as a local temporal smoother to capture short-term correlations, extracting features vectors in $\mathbb{R}^{T \times d_2}$ ($d_2 < d_1$). The sequence is then fed into a unified, single-layer Gated Recurrent Unit [1] to capture long-range temporal dependencies and global scene dynamics. Finally, the temporal hidden states are passed to a shared regression MLP. Consisting of a dropout layer followed by two linear projections and a Sigmoid activation, this head outputs continuous, frame-level probability scores $\hat{y} \in [0, 1]^{T \times |\mathcal{E}|}$ jointly for all possible target micro-events $e \in \mathcal{E}$ ($|\mathcal{E}| = 2$ for jump spotting). Unlike conventional PES frameworks that employ a Softmax activation over a set of target events augmented by an explicit background class, our architecture models each micro-event as an independent continuous regression target, the predicted probabilities across different event classes are, thus, non-mutually exclusive.

Comprehensive ablation studies validating the individual contributions of the Strip Pooling module and the shared GRU are provided in Section 4.5.

3.4 Training Strategy and Loss Formulation

In Precise Event Spotting, target micro-events are inherently sparse. Framing this as a rigid classification problem often leads to severe class imbalance and forces the reliance on heuristic post-processing algorithms like Non-Maximum Suppression to eliminate redundant positive predictions around the event.

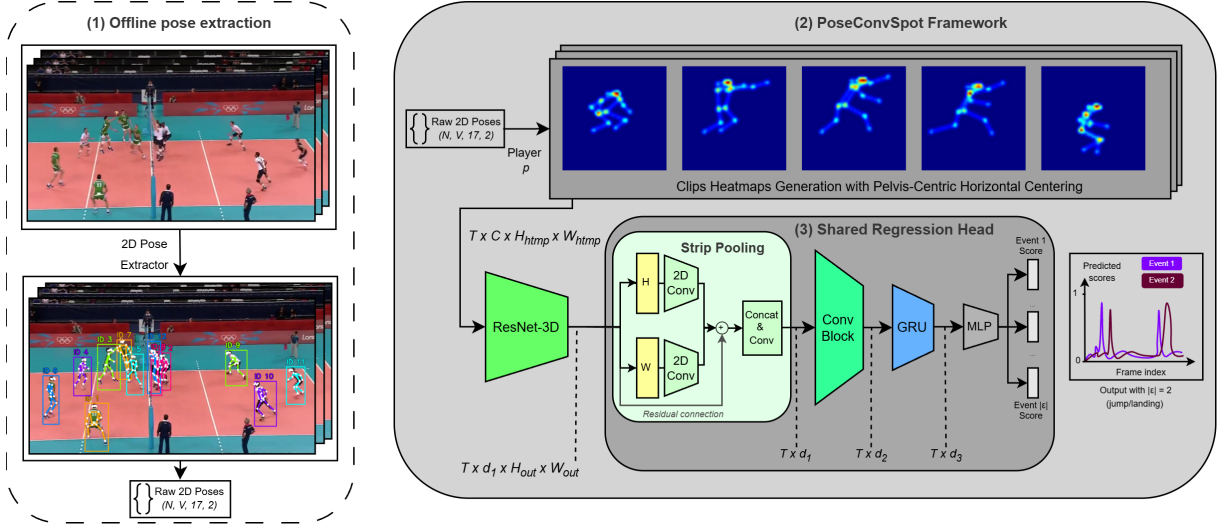


Figure 2: Overview of the PoseConvSpot framework. (1) Offline extraction of 2D joints (N players, V frames, 17 joints) from raw RGB videos. (2) The core pipeline: raw tracklets are transformed into pelvis-aligned pseudo-image heatmaps and processed by a ResNet-3D backbone. A Shared Regression Head, featuring directional Strip Pooling and a GRU, outputs continuous, frame-level probability scores for target events (e.g., jump and landing).

To circumvent this, we formulate PES as a continuous temporal regression task. Similar to Oo et al. [13], the discrete ground-truth timestamps are smoothed using a 1D Gaussian kernel (parametrized by σ), transforming the sparse binary labels into continuous soft targets $y \in [0, 1]$. However, while Oo et al. optimize these targets using a standard MSE which treats all prediction errors uniformly and remains vulnerable to the extreme background-to-event imbalance inherent to PES. To effectively optimize our network against these soft targets while dynamically managing this extreme imbalance, we introduce a Soft Gamma Loss, leveraging the Quality Focal Loss (QFL), originally introduced by Li et al. for dense object detection [11].

For a given frame and specific event class, let y be the soft ground-truth target and $\hat{y}$ be the predicted probability. We first define the standard Binary Cross Entropy for continuous targets:

$$L_{BCE}(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (1)$$

Following the QFL formulation, we apply a dynamic scaling factor based on the absolute prediction error to prevent the vast amount of easy predictions from overwhelming the gradient. In addition, we introduce a balancing factor α to explicitly balance the importance of event peaks versus background. The focal weight w_{focal} and the balancing weight w_{α} are defined as:

$$w_{focal} = |y - \hat{y}|^{\gamma} \quad (2)$$

$$w_{\alpha} = y\alpha + (1 - y)(1 - \alpha) \quad (3)$$

where γ acts as a focusing parameter that down-weights easy examples. We set $\gamma = 2.0$ and $\alpha = 0.25$. The final objective function is computed by averaging across all T frames and $|\mathcal{E}|$ event classes:

$$\mathcal{L} = \frac{1}{T|\mathcal{E}|} \sum_{t=1}^T \sum_{e=1}^{|\mathcal{E}|} w_{\alpha}^{(t,e)} w_{focal}^{(t,e)} L_{BCE}(y_{t,e}, \hat{y}_{t,e}) \quad (4)$$

Table 1: Data statistics (Num of Tracklets | Jump / Landing events) across the three evaluation protocols.

Protocol	Train	Val	Test
Standard	1,448 1084 / 1090	610 486 / 490	648 445 / 443
Data _{0.5}	730 523 / 526	300 240 / 242	648 445 / 443
Data _{0.1}	150 133 / 135	58 63 / 63	648 445 / 443
LOAO	360 296 / 299	80 49 / 49	70 38 / 38

By directly optimizing this continuous formulation, PoseConvSpot learns to predict well-defined probability curves that mimic the shape of the predefined Gaussian targets. Consequently, at inference time, localizing the micro-events simply requires extracting local maxima from the 1D temporal output. Because the model is strictly penalized for outputting high probabilities outside the explicit bounds of the Gaussian distribution, this approach naturally outputs isolated peaks, completely eliminating the need for NMS or any hyperparameter-sensitive deduplication step.

4 Experiments

4.1 Evaluation Setup

Datasets. We evaluate our framework on the *SportsJumpMotion* dataset [13], which provides tracklet-level, frame-accurate annotations for *Jumping* and *Landing* events in professional basketball and volleyball broadcasts. To rigorously assess the capabilities of our model, we employ three distinct evaluation protocols (summarized in Table 1):

- **Standard Benchmark:** Follows the original data split proposed by Yin et al. [13].
- **Cross-Venue Protocol (Leave-One-Arena-Out):** Evaluates the models' immunity to severe visual domain shifts

Table 2: Comparison with state-of-the-art methods on the SportsJumpMotion Standard Benchmark. We report the mAP at various tolerances δ and the F1-Score. For the strongest methods, we report the mean and standard deviation $\pm$ across 3 random seeds. Best results are in bold.

Method	Modality	mAP (%)				F1-Score ($\delta = 3$)
		$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$	
E2E-Spot [7]	RGB	54.31	74.84	80.55	81.86	84.27
T-DEED _{200MF} [23]	RGB	82.67 ± 0.70	91.68 ± 1.00	94.41 ± 0.40	94.95 ± 0.20	91.52 ± 0.81
T-DEED _{800MF} [23]	RGB	84.52 ± 1.70	92.28 ± 0.30	94.66 ± 0.40	95.45 ± 0.40	91.71 ± 0.15
AdaSpot [24]	RGB	82.06	90.11	92.91	93.78	87.28
Ours (Joints)	Pose	78.83 ± 1.62	92.26 ± 0.43	95.91 ± 0.40	96.79 ± 0.12	91.62 ± 0.31
Ours (Limbs)	Pose	79.09 ± 1.06	93.20 ± 0.64	96.05 ± 0.66	96.80 ± 0.50	92.32 ± 0.18
Ours (Full)	Pose	79.08 ± 0.34	93.82 ± 0.20	96.88 ± 0.32	97.70 ± 0.32	93.16 ± 0.52

(e.g., varying court colors, lighting conditions, and camera angles). The data is partitioned such that models are trained and validated on a subset of basketball arenas, and tested on entirely unseen courts. Visual samples of the respective training, validation, and test arenas are provided in Supplementary Material E.

- **Data-Constrained Protocol:** Assesses data efficiency by artificially constraining the standard training and validation sets to 50% and 10% of their original broadcast videos, while retaining the full standard test set. The sampling is stratified by sport; for instance, the Data_{0.1} training split randomly samples 10 of the 99 training volleyball videos and 3 of the 26 training basketball videos. To ensure statistical consistency, the videos sampled for the Data_{0.1} split form a strict subset of the Data_{0.5} split.

Evaluation Metrics and Baselines. Following standard PES protocols [2], we evaluate performance using mean Average Precision (mAP) at strict temporal tolerances $\delta \in \{1, 2, 3, 4\}$ frames, alongside the optimal F1-Score at $\delta = 3$.

We benchmark against three state-of-the-art video-based PES methods: E2E-Spot [7], T-DEED [23], and AdaSpot [24], trained using their official codebases. During inference, all methods employ a 50% overlapping sliding window and Test-Time Augmentation (horizontal flipping). Crucially, for fair comparison, while RGB baselines require Soft-NMS for peak refinement, PoseConvSpot is evaluated natively *without* any post-processing (see Section 4.5). Top results report the mean and standard deviation across three random seeds.

Comparison with [13]: The dataset creators [13] employ a non-standard mAP computation. To ensure alignment with the broader PES literature [2], our main experiments use the official metric. A direct evaluation using their custom metric, along with detailed notes regarding the reproducibility of their original checkpoints, is provided in Supplementary Material A.

Implementation Details. All experiments were conducted on an NVIDIA GeForce RTX 4070 Ti SUPER GPU. If a player leaves the camera field-of-view, missing frames are zero-padded so all tracklets match the video length. This identical strategy is applied to both our framework and the RGB baselines. Regarding spatial resolution, input frames for the RGB baselines are resized to 224×224 pixels.

Conversely, PoseConvSpot computes pseudo-image heatmaps at a lower resolution of 48×48 pixels.

For our method, 2D poses are extracted offline using an off-the-shelf, COCO pretrained ViTPose [27]. Video sequences are partitioned into $T = 100$ frames windows. The network is trained for 80 epochs using the AdamW optimizer, with an initial learning rate of 1×10^{-4} decayed via a cosine annealing schedule following a 10-epoch linear warmup. The Gaussian smoothing window for the soft targets uses a standard deviation of $\sigma = 1.5$ frames. Additional details are provided in Supplementary Material B.

4.2 Comparison to State-of-the-Art

We compare our proposed PoseConvSpot against recent state-of-the-art video-based Precise Event Spotting frameworks. The results on the standard SportsJumpMotion benchmark are summarized in Table 2. To ensure statistical reliability, the mAP and standard deviations for the most competitive methods are averaged across three distinct random seeds.

For all standard operational tolerances ($\delta \geq 2$), our PoseConvSpot framework outperforms all RGB baselines. Interestingly, our method relies exclusively on sparse skeletal representations without accessing any raw visual cues or background context, yet it achieves exceptional spotting precision. At $\delta = 3$, our full model (Joints + Limbs) surpasses the second-best model (T-DEED_{800MF}) by an absolute margin of +2.22%. Furthermore, it outperforms recent architectures like AdaSpot, whose adaptive spatial attention mechanisms struggle to extract meaningful features from pre-cropped player tracklets. This limitation is expected, as such AdaSpot redundancy-aware modules are explicitly designed to leverage full, high-resolution visual contexts to accurately pinpoint discriminative cues.

At the strictest temporal tolerance ($\delta = 1$), our model performs worse than the RGB-based method T-DEED_{800MF} which achieves the highest performance (84.52%). This behavior is expected for a pose-only architecture. By intentionally discarding fine-grained visual details, such as the precise pixel where a shoe makes contact with the floor, our method relies entirely on the skeleton’s kinematics. Consequently, pinpointing the exact frame of an event often suffers from action-dependent ambiguity rather. The difficulty of resolving this ambiguity varies significantly depending on the nature of the jump. For example, a high-intensity volleyball spike typically

Table 3: Evaluation of Unseen Domain Generalization (LOAO) and Data-Efficiency. We report the mAP across the restricted protocols, with all results averaged over 3 random seeds. Best results per metric and per protocol are in bold.

Protocol	Method	mAP@ δ (%)			
		$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$
LOAO	T-DEED _{200MF}	58.11	74.50	76.35	76.93
	T-DEED _{800MF}	61.79	72.61	73.32	73.75
	Ours (Joints)	65.69	83.40	89.57	90.30
	Ours (Limbs)	66.31	82.83	87.95	90.85
	Ours (J + L)	64.58	81.91	90.40	91.23
Data _{0.5}	T-DEED _{200MF}	79.59	88.39	91.40	92.09
	T-DEED _{800MF}	79.82	88.85	91.70	92.51
	Ours (Joints)	76.46	91.65	95.43	96.30
	Ours (Limbs)	75.89	91.97	95.45	96.48
	Ours (J + L)	76.31	91.71	95.08	96.09
Data _{0.1}	T-DEED _{200MF}	59.07	70.87	76.17	77.47
	T-DEED _{800MF}	62.21	75.12	79.29	81.23
	Ours (Joints)	66.29	86.04	91.33	92.98
	Ours (Limbs)	66.03	86.59	91.81	93.63
	Ours (J + L)	67.57	87.32	91.76	93.68

ends with a pronounced landing phase where the player visibly regains balance over several frames, making the event kinematically distinct. In contrast, a low-elevation basketball jump shot features a much more subtle transition. Without RGB context, distinguishing the final airborne frames from the exact moment of ground contact is highly challenging, particularly if the player remains relatively static upon landing.

4.3 Unseen Domain Generalization

In sports broadcasting, pixel-level features are highly susceptible to domain shifts, such as varying court colors, lighting variations, and novel camera angles. To evaluate robustness against these environmental changes, we test the top-performing baselines (T-DEED) and our method under the Cross-Venue (LOAO) protocol.

As reported in Table 3, the RGB-based models suffer a significant performance drop when deployed on unseen arenas. Interestingly, the model with the larger feature extractors (T-DEED_{800MF}) performs worse than its lighter counterpart (T-DEED_{200MF}) at several tolerances ($\delta \geq 2$). As it is a substantially larger model with much higher capacity, it is not surprising that it severely overfits to the background textures and specific camera perspectives seen during training.

Furthermore, it is essential to disentangle the impact of this domain shift from mere data scarcity, as the LOAO protocol inherently reduces the size of the training set. However, a comparison with our data-efficiency experiments (Section 4.4) reveals that the performance degradation in LOAO is significantly more severe than

what data starvation alone would cause. For instance, T-DEED_{800MF} achieves a lower mAP in LOAO (73.32% at $\delta = 3$) than in the extreme Data_{0.1} split (79.29%), even though the LOAO split contains more than twice as many training events. This isolates the pixel-level domain shift as the primary culprit for the RGB model’s collapse.

In contrast, PoseConvSpot retains a highly competitive performance. By relying exclusively on geometric priors, our method effectively neutralizes environmental changes. At $\delta = 3$, our full model (Joints + Limbs) achieves an impressive 90.40% mAP, outperforming the most resilient RGB baseline by an absolute margin of +14.05%. This confirms that isolating the athlete’s topology guarantees much more reliable generalization in real-world broadcast scenarios.

4.4 Data Efficiency

To assess how our framework performs under strict data scarcity independently of domain shifts, we evaluate the models using 50% (Data_{0.5}) and 10% (Data_{0.1}) of the standard training videos.

As the volume of training data decreases, the performance of traditional RGB baselines degrades steadily. To generalize effectively, end-to-end vision models must implicitly learn to disentangle the action of interest from complex environmental variables. This heavy burden of simultaneously learning spatial invariance and temporal dynamics demands massive amounts of varied training data. Consequently, when restricted to just 10% of the dataset, the strongest baseline (T-DEED_{800MF}) drops significantly to 79.29% at $\delta = 3$.

Conversely, PoseConvSpot demonstrates exceptional data efficiency. Under the exact same severe data constraints, our full model maintains a high mAP of 91.76% (+12.47% absolute improvement over the baseline). Even at the strictest tolerance ($\delta = 1$), our model outperforms the RGB equivalent (67.57% vs 62.21%).

This remarkable resilience stems from the highly distilled nature of skeletal heatmaps. Because the pose extractor acts as a robust informational filter that naturally eliminates background clutter and irrelevant visual noise, our spatio-temporal regression head is entirely relieved from the burden of learning spatial invariance from scratch. Instead, it focuses exclusively on the core biomechanical trajectories of the actions, allowing the network to converge successfully and maintain high precision with drastically fewer training examples.

4.5 Ablation Study

To validate our architectural choices, we conduct an ablation study on the standard evaluation protocol using our full model (Joints + Limbs) on a fixed deterministic seed.

Architecture & Loss Formulation. Table 4 summarizes the impact of our core components. Replacing Strip Pooling with Global Average Pooling decreases mAP (−0.88% at $\delta = 3$), confirming that isotropic pooling dilutes directional biomechanical signals like vertical jumps. Removing the GRU layer causes a severe degradation (−3.26% at $\delta = 1$), demonstrating the necessity of long-range temporal memory over static frame-level features. Additionally, our Soft Gamma Loss shows superior results over the W-MSE, proving its ability to natively handle extreme background-to-event imbalance.

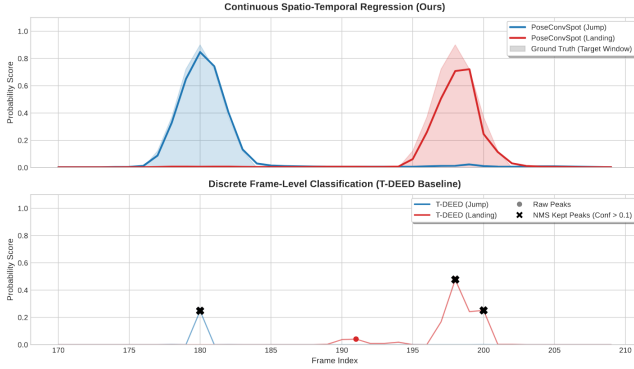
Impact of Post-Processing. Table 5 and Figure 3 highlight the robustness of our continuous regression formulation. Standard RGB architectures rely on discrete classification paradigms that inherently output fragmented, low-confidence probability scores (often

Table 4: Ablation study on the core architecture and training pipeline of PoseConvSpot.

Model Configuration	mAP@ δ (%)			
	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$
PoseConvSpot (J+L)	79.22	93.99	97.14	98.07
<i>Architecture Variations:</i>				
w/o Strip Pooling (use GAP)	78.50	93.07	96.26	97.10
w/o GRU (use MLP)	75.96	92.34	95.01	95.98
<i>Loss & Optimization:</i>				
w/o Soft Gamma (use W-MSE)	78.11	92.79	95.93	96.65

Table 5: Impact of heuristic post-processing on model performance. Unlike RGB baselines, PoseConvSpot naturally outputs isolated peaks.

Method	Post-Processing	mAP@ δ (%)			
		$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$
T-DEED _{200MF}	Soft-NMS	82.67	91.68	94.41	94.95
T-DEED _{200MF}	None	79.35	86.83	89.25	89.70
T-DEED _{800MF}	Soft-NMS	84.52	92.28	94.66	95.45
T-DEED _{800MF}	None	80.45	86.97	89.07	89.71
PoseConvSpot	None	79.22	93.99	97.14	98.07

**Figure 3: Probability outputs for consecutive Jump/Landing events. Top: PoseConvSpot generates smooth, high-confidence curves matching the Gaussian ground truth, requiring zero post-processing. Bottom: The RGB baseline yields fragmented signals. Its heuristic Soft-NMS fails to filter it, notably leaving multiple conflicting peaks during the second landing (frames 190–205).**

peaking below 0.5). As illustrated in Figure 3, without Soft-NMS post-processing to aggressively filter these redundant peaks, models like T-DEED suffer an important performance collapse. Conversely, by directly optimizing temporal consistency via the Soft Gamma Loss, PoseConvSpot inherently outputs smooth, high-confidence peaks that correctly mimic the ground truth, completely bypassing the need for post-processing while achieving superior accuracy.

5 Conclusion and Future Work

In this paper, we introduced **PoseConvSpot**, a pose-only spatio-temporal regression framework designed to tackle the inherent limitations of dense RGB architectures in Precise Event Spotting. By transforming raw skeletal sequences into pelvis-aligned, pseudo-image heatmaps, our approach effectively isolates the core biomechanical kinematics of individual athletes. This design intrinsically shields the network from severe pixel-level domain shifts caused by varying backgrounds and camera angles, while naturally bypassing the spatial occlusion issues that severely hinder RGB models during simultaneous multi-player events. Extensive experiments on the SportsJumpMotion dataset demonstrate that PoseConvSpot not only establishes a new state-of-the-art across standard operational tolerances for jump spotting but also exhibits unprecedented robustness in cross-venue generalization and exceptional data efficiency.

Future Work. While PoseConvSpot succeeds in kinematic-driven spotting, it inherently discards the external visual context required for events defined by human-object interactions (e.g., kicking a ball). Consequently, future work will aim to evaluate and extend our framework across a broader spectrum of sports where events cannot be recognized solely through skeletal data. To handle such complex triggers while preserving the advantages of our approach, a promising direction is to integrate task-related feature maps (e.g., ball position heatmaps) as supplementary input channels of the ResNet, effectively fusing object-centric and pose-centric representations. Furthermore, because our end-to-end efficiency currently relies on pre-requisite pose extraction, future studies will investigate the framework’s robustness to pose estimation noise, specifically evaluating its performance when paired with lighter, real-time pose extractors.

References

- [1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv:1412.3555 [cs.NE]* <https://arxiv.org/abs/1412.3555>
- [2] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. 2021. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4508–4519.
- [3] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. 2024. COMEDIAN: Self-supervised learning and knowledge distillation for action spotting using transformers. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*. 530–540.
- [4] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2969–2978.
- [5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. *arXiv:1812.03982 [cs.CV]* <https://arxiv.org/abs/1812.03982>
- [6] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1711–1721.
- [7] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. 2022. Spotting temporally precise, fine-grained events in video. In *European Conference on Computer Vision*. Springer, 33–51.
- [8] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. 2020. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4003–4012.
- [9] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs.LG]* <https://arxiv.org/abs/1502.03167>
- [10] Christoff Kügler, Patrick Trollmann, Florian Wolling, and Florian Michahelles. 2024. JumpMetric: Assessment of Fiducial Positions for Vertical Jump Height Estimation from Depth Cameras and Wearable Sensors. In *Proceedings of the 23rd International Conference on Mobile and Ubiquitous Multimedia (MUM '24)*. Association for Computing Machinery, New York, NY, USA, 419–431. doi:10.1145/3701571.3701607
- [11] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. 2020. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 21002–21012. https://proceedings.neurips.cc/paper_files/paper/2020/file/f0bda020d2470f2e74990a07a607ebd9-Paper.pdf
- [12] Mengyuan Liu, Hong Liu, Qianshuo Hu, Bin Ren, Junsong Yuan, Jiaying Lin, and Jiajun Wen. 2025. 3d skeleton-based action recognition: A review. *arXiv preprint arXiv:2506.00915* (2025).
- [13] Yin May Oo, Ankhzaya Jamsrandorj, Yanyi Chao, Hoang Quoc Nguyen, Yewon Hwang, Kyung-Ryoul Mun, and Jinwook Kim. 2025. Jump-Aware: Player Position Rectification and Identification in Dynamic Sports Using Jump Event Spotting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 5925–5934. doi:10.1109/CVPRW67362.2025.00591
- [14] Soumyadip Roy, Chaitanya Roygaga, Nathaniel Blanchard, and Aparna Bharati. 2022. A Computer Vision Method for Estimating Velocity from Jumps. *arXiv preprint arXiv:2212.04665* (2022).
- [15] Sanchayan Santra, Vishal Chudasama, Pankaj Wasnik, and Vineeth N Balasubramanian. 2025. Precise Event Spotting in Sports Videos: Solving Long-Range Dependency and Class Imbalance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3163–3172.
- [16] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2616–2625.
- [17] Srishti U Sharma, Srikrishnan Divakaran, Tolga Kaya, and Mehul S Raval. 2025. A Computer Vision Framework on Biomechanical Analysis of Jump Landings. In *Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing (ICVGIP '24)*. Association for Computing Machinery, New York, NY, USA, Article 9, 9 pages. doi:10.1145/3702250.3702259
- [18] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. 2023. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18857–18866.
- [19] Zhou Shuchang. 2022. A survey on human action recognition. *arXiv preprint arXiv:2301.06082* (2022).
- [20] Joao VB Soares, Avijit Shah, and Topojoy Biswas. 2022. Temporally precise action spotting in soccer videos using dense detection anchors. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2796–2800.
- [21] Kim Hoang Tran, Phuc Vuong Do, Ngoc Quoc Ly, and Ngan Le. 2024. Unifying global and local scene entities modelling for precise action spotting. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
- [22] Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. 2023. Astra: An action spotting transformer for soccer videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*. 93–102.
- [23] Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. 2024. T-deed: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3410–3419.
- [24] Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. 2026. AdaSpot: Spend Resolution Where It Matters for Precise Event Spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24010–24020.
- [25] Huifen Xia and Yongzhao Zhan. 2020. A Survey on Temporal Action Localization. *IEEE Access* 8 (2020), 70477–70487. doi:10.1109/ACCESS.2020.2986861
- [26] Hao Xu, Arbind Agrahari Baniya, Sam Well, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. 2025. Deep Learning for Sports Video Event Detection: Tasks, Datasets, Methods, and Challenges. *arXiv preprint arXiv:2505.03991* (2025).
- [27] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems* 35 (2022), 38571–38584.
- [28] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [29] Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*. Springer, 492–510.