

Supplementary Material

PoseConvSpot: Overcoming Domain Shifts in Precise Event Spotting

Noâm Boussouf, Adrien Maglo, Mohamed Chaouch, Paul Pham Dang, and Hideo Saito

A Analysis of Custom Evaluation Metrics and Baseline Reproducibility

As noted in Section 4.1 of the main manuscript, our primary benchmark adopts the standardized SoccerNet protocol [2] to ensure a mathematically rigorous, confidence-agnostic evaluation based on the global Area Under the Curve (AUC). For transparency, we outline two fundamental mathematical deviations in the custom evaluation script introduced by Oo et al. [13] that precluded its use as our primary metric. First, rather than computing the global AUC across all confidence thresholds, their protocol applies a **fixed** threshold ($p \geq 0.35$) prior to evaluation and calculates an **arithmetic** mean of discrete precision values, effectively reducing the metric to a single-operating-point precision. Second, their script performs **macro-averaging** by computing scores independently per video before averaging them. In sparse tasks such as jump spotting, this approach disproportionately weights videos with few or no events, mathematically distorting the final metric compared to a standard global (micro) evaluation.

A Note on Reproducibility. Furthermore, we reached out to the authors of [13], who graciously provided their codebase and dataset. While we greatly appreciate their initial assistance, we were ultimately unable to reproduce the exact performance figures reported in their paper. We observed that the released checkpoints appeared to differ from the models explicitly trained on the Sport-sJumpMotion dataset. Therefore, to ensure absolute fairness and transparency, the baseline results reported in Table S1 for Oo et al. are strictly the ones reported in their original article, while PoseConvSpot is evaluated using their provided custom evaluation script.

Table S1: Comparison of overall performance on the Sport-sJumpMotion dataset evaluated under the custom protocol by Oo et al. [13]. Metrics marked with [†] are reported from the original paper. PoseConvSpot is evaluated using their script to allow a direct comparison.

Method	Backbone	mAP _{Oo et al.}		
		$\delta = 1$	$\delta = 2$	$\delta = 3$
E2E-Spot [7]	RegNet-Y	89.39 [†]	92.99 [†]	94.05 [†]
T-DEED [23]	RegNet-Y	89.36 [†]	92.90 [†]	94.48 [†]
Oo et al. [13]	ViT-B	90.80[†]	94.41 [†]	95.33 [†]
Ours	ResNet3D (Pose)	90.20	94.94	95.79

To enable a direct comparison, Table S1 presents our results evaluated strictly under the custom protocol of Oo et al. [13]. The threshold for keeping a detection is fixed at 0.35. Despite the complete absence of dense visual features, PoseConvSpot remains highly competitive with the heavy ViT-B baseline. Notably, our pose-only framework surpasses their architecture in overall mAP at standard

temporal tolerances ($\delta = 2$ and $\delta = 3$). While Oo et al. retain a marginal advantage at the strictest tolerance ($\delta = 1$), our approach establishes superior localization robustness across broader temporal windows, demonstrating that PoseConvSpot matches or exceeds state-of-the-art RGB architectures even under a non-standard metric optimized for predefined operating points.

B Implementation details

To ensure reproducibility, in this section we provide additional implementation details for PoseConvSpot.

Network Architecture and Heatmap Generation Our framework utilizes a ResNet3D-50 backbone to process the spatial features of the pseudo-image heatmaps. To generate these 48×48 inputs, 2D joint coordinates are mapped onto a grid where a 2D Gaussian distribution is applied at each joint location with a standard deviation of $\sigma = 0.8$. To account for bounding box variations, a spatial padding factor is applied. During inference, this bounding box padding is fixed at +0.2 (relative to the player’s scale). To mitigate overfitting, dropout is applied with a probability of 0.4.

Data Sampling and Training Configuration Training clips are generated using a sliding window approach with a temporal stride of 60 frames (e.g., if the first clip covers frames 0 to 100, the second covers frames 60 to 160). Clips containing more than 60% zero-padded frames (indicating the player is absent from the camera field-of-view) are discarded during training.

The network is trained with a batch size of 16 and a weight decay of 5×10^{-3} . Empirically, we observed that while the exact value of the weight decay had a marginal impact on performance when optimizing our Soft Gamma loss, it remained crucial for achieving the best possible results when training with the standard Mean Squared Error (MSE) loss. Furthermore, to prevent the network from being disproportionately penalized for highly confident predictions that fall marginally short of a perfect score, we multiply the soft Gaussian targets by a factor of 0.9 (setting 0.9 as the maximum target value for a jump frame). All reported results are averaged across three distinct training seeds (25, 26, and 27).

Data Augmentation: The data augmentations methods parameters applied during training are detailed as follows:

- **Temporal Shifting:** Training clips are dynamically sampled by extracting $T = 100$ consecutive frames from an extended 120-frame window, effectively applying a random temporal shift of ± 10 frames.
- **Spatial Padding:** While fixed at +0.2 during inference, the padding factor is sampled uniformly from the interval $[0.1, 0.3]$ during training.
- **Joint-level Perturbations:** To simulate pose estimation inaccuracies, random Gaussian noise with a standard deviation of 0.5 is added independently to the coordinates of each joint with a probability of 0.5. Additionally, up to 2 joints per frame are randomly dropped with a probability of 0.3.



Figure S1: Visual samples of split arenas in the Leave-One-Arena-Out evaluation protocol. Columns represent distinct arenas to illustrate the domain shift encountered during evaluation under this protocol.

- Geometric Transformations: Horizontal flipping is applied to the input heatmaps with a probability of 0.5.

C Data augmentation

Table S2 isolates the impact of our skeletal augmentations. Removing all specialized augmentations results in a substantial -2.57% mAP drop at $\delta = 3$. Ablating them individually reveals that each augmentation uniquely contributes to the model’s robustness.

Table S2: Ablation of specialized skeletal data augmentations. The baseline always includes standard horizontal flipping.

Applied Augmentations				mAP@ δ (%)			
Temp. Shift	Drop	Noise	Pad	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$
×	×	×	×	76.13	91.00	94.57	95.81
×	✓	✓	✓	78.04	91.35	95.40	96.23
✓	×	✓	✓	77.07	92.31	96.49	97.61
✓	✓	×	✓	77.78	93.11	96.21	97.53
✓	✓	✓	×	78.12	92.08	95.80	96.93
✓	✓	✓	✓	79.22	93.99	97.14	98.07

D Overlap Ratio at Inference

During inference, our sliding window approach relies on an overlap ratio to process continuous video streams. A higher overlap means that consecutive clips share more frames, which allows for smoother predictions by averaging the scores across overlapping regions, but also increases the overall computational cost.

Table S3 presents an ablation study evaluating the impact of this temporal overlap ratio on the performance of our Joints+Limb model. We compare overlap ratios ranging from 0% (no overlap, contiguous clips) to 90% (highly dense sliding window).

As shown in Table S3, the temporal overlap ratio during sliding-window inference has a marginal impact on precision. Even with

Table S3: Impact of the inference overlap ratio on the performance of the Joints+Limb model. Metrics are reported as percentages.

Overlap Ratio	mAP@1	mAP@2	mAP@3	mAP@4
0%	80.13	93.06	96.73	97.66
25%	80.11	93.49	97.26	98.08
50%	79.22	93.99	97.14	98.07
75%	80.56	93.39	97.16	98.13
90%	80.31	93.95	97.49	98.37

0% overlap, PoseConvSpot remains highly stable (96.73% mAP@3), proving its intrinsic temporal robustness. Raising the overlap to 90% yields only minor gains (97.49% mAP@3), but significantly increases inference latency due to dense redundant forward passes. To maintain a standard, fair comparison with existing baselines, we employ the standard 50% overlap throughout our main experiments.

E Leave-One-Arena-Out Visual Samples

Figure S1 displays representative frames from our training, validation, and testing splits. To illustrate the domain shift encountered during Leave-One-Arena-Out (LOAO) evaluation, each column showcases distinct arenas, highlighting the variability in camera angles and court layouts.

F Model Complexity and Parameter Count

Table S4 compares the trainable parameter count of PoseConvSpot variants against state-of-the-art video-based PES methods. Our pose-only architecture introduces significantly lower overhead than heavy RGB pipelines like T-DEED, which rely on massive backbones to process high-resolution video streams. Instead, PoseConvSpot operates on compact 48×48 pseudo-image heatmaps via

a standard 3D convolutional backbone. Assuming 2D poses are pre-extracted—a common prerequisite for downstream tasks like 3D game-state reconstruction—this reduced complexity and the absence of heuristic post-processing make our framework highly efficient for practical deployment without sacrificing frame-level accuracy.

Table S4: Comparison of trainable parameters (in millions) between state-of-the-art PES architectures and our PoseConvSpot variants.

Method	Backbone	Parameters (M)
E2E-Spot [7]	RegNet-Y	12.61
T-DEED _{200MF} [23]	RegNet-Y	12.31
T-DEED _{800MF} [23]	RegNet-Y	46.22
AdaSpot [24]	RegNet-Y	10.63
Oo et al. [13]	ViT-B	87.84
PoseConvSpot (Joints)	ResNet3D-50	4.54
PoseConvSpot (Limbs)	ResNet3D-50	4.54
PoseConvSpot (Full)	ResNet3D-50	4.57