

Spatiotemporal Event Spotting via 3D Heatmaps with Dynamically Shifted Gaussian Kernels

Ankhzaya Jamsrandorj^{*1}

ankhzaya@kist.re.kr

Vanyi Chao^{*1,2}

vanyi@kist.re.kr

Hoang Quoc Nguyen¹

523503@kist.re.kr

Yin May Oo^{1,2}

yinmay@kist.re.kr

Muhammad Amrulloh Robbani^{1,2}

amrullohrobbani@kist.re.kr

Yewon Hwang¹

ywhwang@kist.re.kr

Kyung-Ryoul Mun¹

krmoon02@kist.re.kr

Jinwook Kim¹

jinwook.kim21@gmail.com

¹ Intelligence and Interaction Research Center

Korea Institute of Science and Technology

Seoul, Republic of Korea

² AI-Robotics, KIST School

University of Science and Technology,
Seoul, Republic of Korea

^{*}These authors contributed equally

Abstract

Spatiotemporal event spotting in ball sports is essential for understanding complex game dynamics, requiring both high temporal precision and accurate spatial localization. However, most existing methods focus primarily on temporal localization, often neglecting the spatial dimensions that are crucial for tactical analysis. In this study, we propose a novel representation, 3D Heatmaps with Dynamically Shifted Gaussian Kernels, specifically designed to enable comprehensive spatiotemporal event spotting. To overcome current limitations, we introduce the Volleyball Nations League (VNL) dataset, which includes detailed annotations for eight key event types, encompassing both temporal and spatial labels. Our approach leverages a modified 3D U-Net architecture that effectively captures spatiotemporal patterns by utilizing our proposed heatmap design. Experimental results show that our method significantly outperforms state-of-the-art techniques in both temporal accuracy and spatial precision on the VNL dataset and a spatially augmented version of the SoccerNet Ball Action Spotting (BAS) dataset. These findings demonstrate the robustness and generalizability of our approach across different ball sports domains.

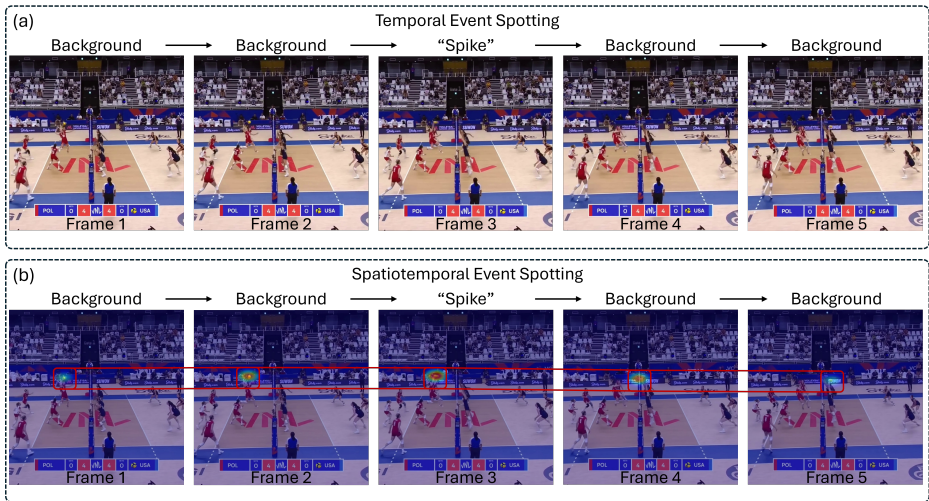


Figure 1: Comparison between Temporal Event Spotting (TES) and Spatiotemporal Event Spotting (STES). (a). TES focuses only on detecting the temporal frame where an event occurs, without considering its spatial location in the scene. (b). STES extends this task by localizing the event both temporally and spatially, identifying not only when but also where the event takes place within the frame.

1 Introduction

Understanding complex actions and key events in sports videos is a fundamental challenge in video understanding. While early research primarily focused on action classification, recent work has advanced to more detailed and demanding tasks such as Temporal Action Localization (TAL), Spatiotemporal Action Localization (STAL), and Temporal Event Spotting (TES). These tasks aim not only to recognize what is happening in a video but also to determine when and where it occurs.

In ball sports such as volleyball, soccer, or basketball, specific events like a spike or goal directly influence the outcome of a game. Accurately spotting these events at the exact moment and location is essential for applications such as tactical analysis, team and player performance evaluation, and highlight generation. However, most existing event spotting works focus solely on identifying when an event occurs (Fig. 1(a)), while neglecting where it happens, a crucial element for a deeper understanding of game dynamics.

To address this limitation, we define and investigate the task of Spatiotemporal Event Spotting (STES), which requires detecting discrete events at both the frame level and their corresponding spatial location within the scene. STES combines the frame-level temporal precision of TES with the spatial awareness of STAL. Unlike STAL, which typically tracks the actor performing an action over multiple frames, STES focuses on spotting a single spatial location at a specific frame where a discrete event occurs, such as ball-ground contact or player-ball interaction (Fig. 1(b)). Although localization is performed in image space, these pixel-level coordinates can be projected onto the court using homography, enabling spatial distribution analysis of events and tactical insights into gameplay.

Despite the importance of STES for sports video analysis, no existing dataset or model explicitly supports the STES task. To this end, we introduce the Volleyball Nations League

(VNL) dataset, consisting of 3,518 rallies and over 23,000 annotated events across eight key categories: *Serve*, *Receive*, *Set*, *Spike*, *Block*, *Dig*, *Pass*, and *Score*. Each event is annotated with its precise temporal frame and the ball-based spatial location. In addition, we enhance the SoccerNet Ball Action Spotting (BAS) dataset by annotating spatial locations for each event, allowing us to evaluate the generalizability of our approach.

To model the STES task, we propose a 3D heatmap representation with dynamically shifted Gaussian kernels, which encodes the likelihood of event occurrences across both spatial and temporal dimensions. For each event, a Gaussian kernel is centered at the annotated pixel location and dynamically shifted across a temporal window to reflect the uncertainty and spread of the event across nearby frames. This formulation creates a dense volumetric heatmap that emphasizes spatiotemporal continuity while preserving localization sharpness. We adapt a simple yet effective 3D U-Net [24] architecture to learn this representation, leveraging its encoder-decoder structure to aggregate contextual features across space and time. Despite its simplicity, our framework achieves state-of-the-art performance on the proposed VNL dataset and the spatially augmented BAS dataset, demonstrating strong generalization across different sports domains.

Our contributions are summarized as follows:

- **Definition of a new task:** We formally define the task of Spatiotemporal Event Spotting (STES), which requires the joint temporal and spatial localization of discrete, frame-specific events in ball sports videos.
- **Creation of a benchmark dataset:** We introduce the VNL dataset, which features professional volleyball match videos with spatiotemporal event annotations, per-frame ball positions, and court homography.
- **Development of an effective framework:** We propose a novel 3D heatmap representation with dynamically shifted Gaussian kernels for spatiotemporal event modeling. Combined with a simple yet effective 3D U-Net backbone, our method achieves state-of-the-art performance while maintaining generalization across sports domains.

2 Related Works

Recent advances in deep learning and video understanding have expanded the field from basic recognition tasks to more complex challenges, including Temporal Action Localization, Spatiotemporal Action Localization, and Event Spotting. These tasks aim not only to classify actions or events in videos, but also to localize them precisely in time and space.

Temporal Action Localization (TAL) aims to identify the start and end times of actions within untrimmed videos. This capability is critical for applications such as sports analytics, surveillance, and content-based video retrieval. TAL methods are typically categorized into two-stage and one-stage approaches. Two-stage methods [8, 14, 17, 21, 22] generate temporal proposals for candidate action segments and then refine these proposals through classification and regression. In contrast, one-stage methods [16, 18, 23] adopt a unified architecture that performs classification and localization simultaneously, allowing for greater efficiency and end-to-end optimization.

Spatiotemporal Action Localization (STAL) extends the TAL approaches by identifying the spatial regions of actions within each frame. This added dimension provides a richer understanding of scene dynamics, particularly in scenarios involving complex interactions

between humans and objects. Traditional STAL methods [10, 9, 20, 23] often relied on region proposal networks and object tracking. More recent approaches [9, 11, 12, 29] leverage attention mechanisms and transformer-based models to jointly capture temporal and spatial dependencies across frames.

Temporal Event Spotting (TES) introduces a more fine-grained task that focuses on identifying the exact frames in which specific events occur. This level of precision is especially valuable in domains like sports, where detecting key moments such as a pass, goal, or spike is crucial. While some TES methods [10, 11, 22, 25] follow a similar structure to TAL by classifying frames as action or background and refining their locations through temporal regression, others employ fully end-to-end architectures [15, 26] that directly predict event occurrences with high temporal resolution.

Spatiotemporal Event Spotting (STES) has recently emerged as a natural extension of TES, requiring the detection of discrete events with both frame-level temporal accuracy and precise spatial localization. While TES focuses on when an event occurs and STAL captures where and over what duration an action unfolds, STES aims to localize events at a specific frame and location within the scene. Such fine-grained localization is essential in sports video analysis, where key moments such as a ball-ground contact or a scoring event must be detected with high temporal and spatial accuracy. Furthermore, the spatial distribution of events on the court provides valuable insight into team strategies and game tactics, allowing deeper analysis of player positioning, attack zones, and defensive formations. However, to the best of our knowledge, no prior work has explicitly addressed STES as a standalone task. Existing datasets and methods either lack spatial annotations, such as SoccerNet [11] for TES, or focus on extended action intervals rather than frame-level events, as seen in AVA [13] for STAL. Our work addresses this gap by formally defining the STES task and introducing a new dataset and model that jointly capture both temporal and spatial aspects of events.

3 Materials and Methods

3.1 Dataset

Due to the lack of a comprehensive and high-quality dataset for the STES task, we curated and annotated a new dataset of volleyball videos. The full match recordings were sourced from YouTube with permission from the official Volleyball Nations League (VNL) channel. The VNL is an annual international competition organized by the Fédération Internationale de Volleyball (FIVB) to promote volleyball globally.

We collected 24 full-match videos (12 men’s and 12 women’s) from the 2022 and 2023 seasons. The videos were recorded in full HD and downsampled to 25 frames per second. A sports field registration model [19] was employed to segment rallies based on detected court keypoints, resulting in 3,518 rally clips. Each rally comprises between 48 and 1,266 frames,

Number of Games	Number of Rallies	Duration (min)	Number of Instances							
			Serve	Receive	Set	Spike	Block	Dig	Pass	Score
24 (12 female & 12 male)	3518	579.5	3497	3008	4394	4634	2182	2775	488	2660

Table 1: Statistics of the VNL dataset.

depending on the number of events it contains. The number of events per rally ranges from 1 to 42. Table 1 summarizes key statistics of the VNL dataset, including event instance counts.

We defined eight key events: *Serve*, *Receive*, *Set*, *Spike*, *Block*, *Dig*, *Pass*, and *Score*, to comprehensively describe rally dynamics. Definitions of these events are provided in Table 2. Since each event involves ball-player or ball-ground interactions, spatial locations were annotated based on the ball’s position at the time of the event.

Spatiotemporal event annotations were generated using a semi-automatic pipeline: pseudo-labels were first predicted by a model pretrained on a private dataset, then manually refined using a dedicated annotation tool. The final VNL dataset includes video frames with extensive annotations, such as spatiotemporal event labels, per-frame ball positions estimated by a ball tracker [2], and court homography data.

To evaluate the generalizability of our approach, we also employed the SoccerNet Ball Action Spotting (BAS) dataset [5], originally designed for temporal event spotting (TES). We extended it by annotating spatial locations for each event and refining the original temporal annotations to correct some misalignments. We plan to publicly release the VNL and the spatially augmented BAS datasets to facilitate future research in this domain.

#	Event	Definition
1	Serve	Initiates each rally by hitting the ball over the net into the opponent’s court.
2	Receive	Receiving the opposing team’s serve and passing the ball to the setter.
3	Set	Positioning the ball for an attacker as the second contact.
4	Spike	Jumping and forcefully striking the ball over the net offensively.
5	Block	Front-row players jump to stop the opponent’s attack at the net.
6	Dig	Defensive technique to prevent the ball from touching the court post-attack.
7	Pass	Maneuvering the ball between teammates and over the net, excluding receive or dig.
8	Score	Awarding points when the ball contacts the ground in the opponent’s court.

Table 2: Definitions of the eight annotated event types in the VNL dataset.

3.2 3D Heatmap with Dynamically Shifted Gaussian Kernels

Existing end-to-end event-spotting methods such as E2E-Spot [15] and T-DEED [26] frame the task as classification, using a one-hot vector to indicate events (1) versus background (0). T-DEED further augments this approach with a displacement vector to improve temporal precision. In contrast, our model employs a heatmap-based representation, which has proven effective for precise keypoint localization in applications such as ball detection and tracking [2], field registration [19], and human pose estimation.

We propose a novel spatiotemporal heatmap representation with dynamically shifted Gaussian kernels (see Fig. 2) where each temporal slice is centered not only around a target timestamp but also spatially anchored to the corresponding object position, such as the ball. A 1D Gaussian function is applied along the temporal axis, peaking at the annotated event frame

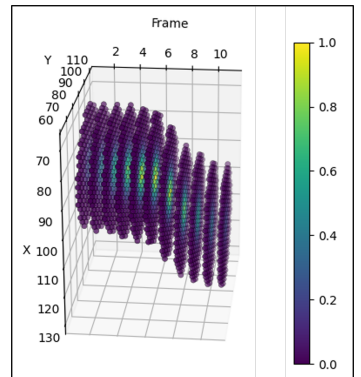


Figure 2: 3D heatmap representation with dynamically shifted Gaussian kernels.

and decaying nonlinearly across time. Simultaneously, each temporal slice in the heatmap volume is associated with a 2D Gaussian kernel spatially centered at the dynamically shifting object location for that frame. This results in a 3D volume in which temporal slices are not aligned to a fixed spatial center but instead shift in space across time, following the object trajectory. Hence, the heatmap preserves motion-aware localization while emphasizing temporal proximity. The representation effectively encodes both when and where events occur, with high resolution in the vicinity of the spatiotemporal center, improving detection of precise, transient sports events like a spike, pass, or shot.

To construct the proposed 3D heatmap, we first define a temporal window neighboring the event frame $t \in \{-\Delta T, -\Delta T + 1, \dots, \Delta T\}$. Within this window, each offset t is used to calculate the 1-D Gaussian temporal weights, expressed as

$$\omega(t) = \exp\left(-\frac{t^2}{2\sigma_t^2}\right), \quad t = -\Delta T, \dots, \Delta T. \quad (1)$$

This weight attains its maximum at $\omega(t=0) = 1$ (at the event frame) and decays symmetrically for earlier or later frames. The 3D heatmap with dynamically shifted Gaussian kernels is then given by

$$H(x, y, t) = \omega(t) \exp\left(-\frac{(x-x_0)^2 + (y-y_0)^2}{2\sigma_s^2}\right)$$

where (x_0, y_0) are the pixel coordinates of the ball in the event frame. Here, σ_s controls the spatial spread and σ_t controls the temporal decay (see Fig. 1 and Fig. 2).

3.3 Model

Following other heatmap-based models, we adopt the U-Net segmentation architecture. However, our task requires a model to be able to process a clipped video and generate the 3D heatmap. To keep the architecture simple yet effective, we employ a 3D version of U-Net [4], adapted from [24].

Fig. 3 shows the full architecture of the 3D U-Net. Given a short video clip $x \in \mathbb{R}^{3 \times T \times H \times W}$, the input is processed through four stages of the encoder. Each encoder block consists of a MaxPool3d layer followed by a double Conv3d block, where each Conv3d block is composed of a GroupNorm layer, a Conv3d layer, and a ReLU activation. Except for the first stage, which does not include a MaxPool3d layer, each subsequent encoder in the downsampling path reduces the temporal length (T), height (H), and width (W) by half, resulting in progressively coarser feature representations. At the end of stage 4, a feature map $x_4 \in \mathbb{R}^{512 \times T/8 \times H/8 \times W/8}$ is obtained, which serves as the bottleneck representation for the decoder path.

In the upsampling path, each decoder stage begins by applying the nearest-neighbor up-sample layer to double the temporal and spatial dimensions of the feature map. The upsampled features are then concatenated with the corresponding encoder features from the skip connections, which helps preserve fine-grained spatiotemporal details. This concatenated tensor is passed through the same double Conv3d block used in the encoder, progressively refining the features and reconstructing the resolution. The final decoder output is passed through a single Conv3d layer with the kernel size of (1,1,1) to generate the 3D heatmap with dynamically shifted Gaussian kernels $\hat{y} \in \mathbb{R}^{C_{\text{out}} \times T \times H \times W}$, where C_{out} denotes the number of events.

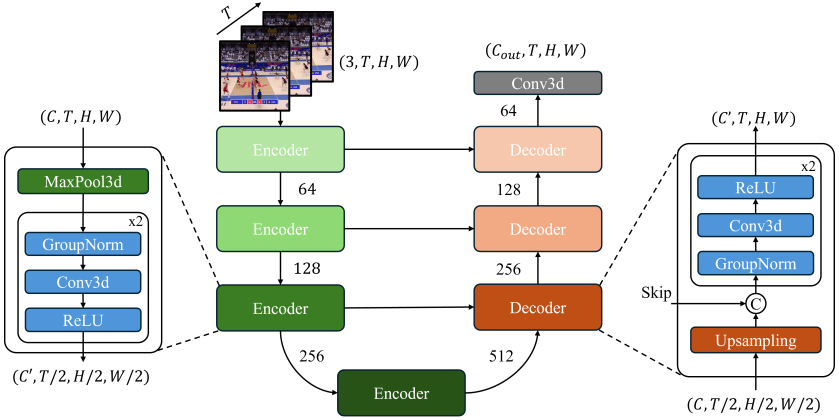


Figure 3: Overview of the proposed 3D U-Net architecture for spatiotemporal event spotting. The network takes a short video clip as input and processes it through a series of encoder and decoder blocks, producing the proposed 3D heatmaps corresponding to the input sequence of T frames. The numbers next to the arrows indicate the output channels of the preceding block and the input channels of the subsequent block.

4 Experiments

4.1 Implementation Details

We split the dataset by game. Of the 24 games, 14 (about 60% of the dataset) were used for training, while the remaining 10 were evenly divided between the validation and test sets. Given a video, we sample short clips of $T = 50$ frames with an overlap of $\frac{2T}{3}$ frames between consecutive clips. For the training set, we sample clips centered around event frames, which reduces the number of samples containing only background. Each input frame is resized to a resolution of 224×224 . For the evaluation and test sets, short clips are uniformly sampled across the entire video. To generate the 3D heatmap with dynamically shifted Gaussian kernels, we set the temporal window $\Delta T = 5$, the temporal standard deviation $\sigma_T = 3$, and the spatial standard deviation $\sigma_S = 5$. To demonstrate the effectiveness of our method, we also train the VNL and SoccerNet BAS datasets using the E2E-Spot [15] and T-DEED [76] models from scratch for comparison.

During training, we apply one random augmentation per sample to enhance generalization. Based on predefined probabilities, we select from geometric augmentations (flip, rotation, perspective, translation, crop), intensity augmentations (brightness, contrast, random noise, Gaussian blur), or no augmentation. Geometric augmentations are applied to both video frames and heatmaps, whereas intensity augmentations affect only the video frames. The model is trained using mean squared error (MSE) loss between the predicted and ground-truth 3D heatmaps. We use the AdamW optimizer with a learning rate of 0.0001 and a weight decay of 0.01, along with the CosineAnnealingWarmupRestarts scheduler. Training is conducted for 50 epochs on a single NVIDIA RTX A6000 GPU.

Model	mAP@2			mAP@4		
	T-DEED	E2E-Spot	Ours	T-DEED	E2E-Spot	Ours
Overall	80.80	85.31	85.63	81.06	86.69	86.14
mAP per class:						
SERVE	90.53	99.80	99.84	90.53	99.80	99.84
RECEIVE	88.85	90.15	90.65	88.85	90.15	90.65
SET	88.11	89.63	89.38	88.30	89.46	89.91
SPIKE	90.63	90.88	90.68	89.95	90.68	90.68
BLOCK	70.37	82.82	85.31	72.30	83.64	85.83
DIG	81.68	85.93	88.16	81.74	86.31	88.66
PASS	54.82	61.10	54.36	55.37	61.10	56.81
SCORE	81.40	86.08	87.00	81.45	86.08	87.00

Table 3: Comparison of temporal event spotting performance on the VNL dataset.

4.2 Evaluation Metrics

Following standard evaluation protocols for event spotting [8, 9, 26], we use mean Average Precision (mAP) to assess our model’s performance. This metric measures the area under the Precision-Recall curve at different temporal and spatial tolerances δ . Specifically, mAP is computed by averaging the Average Precision (AP) across all event classes. For the VNL dataset, we use temporal tolerances $\delta_T = 2, 4$ frames and spatial tolerances $\delta_S = 2, 4, 6, 8$ pixels. For the SoccerNet BAS dataset, we adopt both tight and loose evaluation protocols. The tight setting uses temporal tolerances ranging from 1 to 5 seconds, while the loose setting extends the range from 5 to 60 seconds.

4.3 Results

We evaluated our proposed method on the VNL and the spatially augmented BAS dataset. As shown in Table 3, our model achieves the highest performance in temporal event spotting in 2-frame tolerance, with an mAP of 85.63. At 4-frame tolerance, our model achieves a competitive mAP of 86.14, slightly below E2E-Spot, while still significantly surpassing T-DEED. These results demonstrate our approach’s substantial temporal precision and robustness, especially under stricter evaluation settings. At the class level, our method consis-

Model	Tight Average-mAP			Loose Average-mAP		
	E2E-Spot	T-DEED	Ours	E2E-Spot	T-DEED	Ours
Overall	35.46	36.54	57.25	42.17	44.01	64.02
mAP per class:						
PASS	77.06	81.61	88.58	86.21	86.55	89.01
DRIVE	76.07	79.10	88.35	86.08	85.91	88.78
HEADER	46.24	46.65	81.32	51.66	60.66	89.05
HIGH PASS	51.15	64.83	84.43	57.52	68.26	87.95
OUT	40.23	23.48	69.15	49.47	35.79	82.25
CROSS	49.40	48.14	70.64	53.25	58.08	84.40
THROW IN	0.000	0.000	61.10	0.000	0.000	76.47
SHOT	22.08	32.46	74.63	28.17	51.54	90.38
BALL P. BLOCK	8.790	6.750	68.80	17.51	18.30	79.96
PLAYER S. TACKLE	0.000	3.930	0.000	3.100	8.440	0.000
FREE KICK	54.55	54.55	0.000	72.11	54.55	0.000
GOAL	0.000	0.000	0.000	0.000	0.000	0.000

Table 4: Comparison of temporal event spotting performance on the SoccerNet BAS.

Dataset	mAP			
	@2	@4	@6	@8
VNL	71.10	82.95	88.22	91.16
SoccerNet BAS	48.65	59.10	63.96	65.79

Table 5: Spatial localization performance at varying pixel error thresholds (δ_S) on the VNL and SoccerNet BAS datasets.

tently performs well across most event types, especially in *BLOCK*, *DIG*, and *SCORE*, where precise spatiotemporal reasoning is essential. Although *PASS* remains the most challenging class due to subtle visual cues, our model maintains competitive performance.

To further assess cross-domain generalizability, we evaluate the SoccerNet BAS dataset (Table 4). Under tight and loose conditions, our model achieves the highest average mAP, with 57.25 (tight) and 64.02 (loose), significantly outperforming E2E-Spot and T-DEED. The most significant improvements are observed in visually complex classes such as *HEADER*, *SHOT*, and *CROSS*. Notably, our model is able to detect events such as *THROW IN*, which all baselines fail to identify, by explicitly leveraging spatial cues around the ball that are essential for distinguishing such actions. These gains highlight the robustness of our 3D heatmap formulation in different sports.

However, we observe that our model is sensitive to class imbalance, failing to detect rare event types such as *PLAYER SUCCESSFUL TACKLE* and *FREE KICK*. This suggests that future improvements could incorporate techniques to handle low-frequency events better, such as class reweighting or more data augmentation for the rare classes.

To evaluate our method’s spatial precision, we measure mAP at different pixel thresholds. As shown in Table 5, our method achieves excellent localization performance, especially on the VNL dataset. For example, it obtains 71.1 mAP at 2 pixels, increasing to 91.16 at 8 pixels, indicating accurate frame-level and pixel-level localization. In contrast, the SoccerNet BAS dataset presents more spatial ambiguity due to cluttered scenes and less structured environments. Still, our model performs competitively, reaching 65.79 mAP at 8 pixels, confirming its ability to generalize across sports and video conditions.

We conducted an ablation study to analyze the impact of our 3D heatmap representation with dynamically shifted Gaussian kernels compared to a standard volumetric 3D heatmap, where Gaussian peaks are centered statically at the labeled event locations without adaptation to context. Table 6 shows that the 3D heatmap with a dynamically shifted Gaussian kernel significantly outperforms the standard variant in spatial and temporal event spotting. For spatial event spotting at a 2-pixel threshold, performance improves from 60.73 to 71.1 mAP. In temporal event spotting, performance improves even more drastically: from 60.19 to 80.88 mAP at a 1-frame threshold. These improvements confirm that the proposed 3D heatmap, which effectively emphasizes spatial and temporal regions of interest, helps the model learn sharper and more precise event boundaries. It provides a richer spatiotemporal representation than a uniform 3D volume, leading to consistent gains in both accuracy and robustness.

Heatmap Representation	Spatial mAP (δ_S)				Temporal mAP (δ_T)			
	@2	@4	@6	@8	@1	@2	@3	@4
Standard 3D volumetric heatmap	60.73	73.37	76.13	77.38	60.19	62.15	62.89	64.83
Dynamically Shifted 3D heatmap	71.10	82.95	88.22	91.16	80.88	85.63	85.90	86.14

Table 6: Ablation study comparing Standard 3D volumetric heatmap and 3D heatmap with dynamically shifted Gaussian kernels for spatial and temporal event spotting.

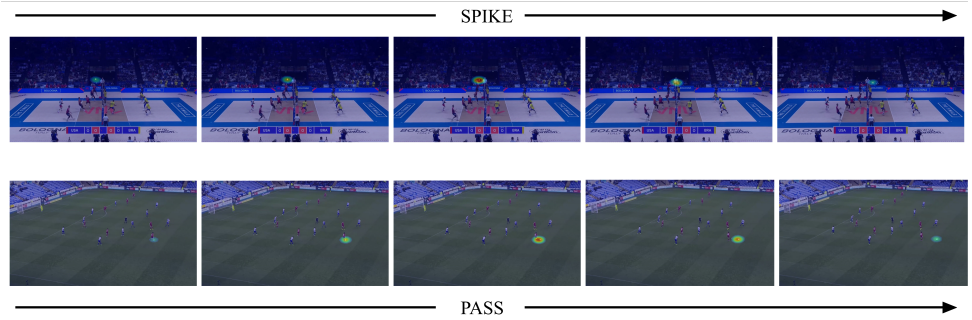


Figure 4: Visualization of predicted spatiotemporal heatmaps overlaid on video frames.

Fig. 4 illustrates the effectiveness of our proposed 3D heatmap with dynamically shifted Gaussian kernels by visualizing predicted spatiotemporal heatmaps overlaid on video frames from the VNL and SoccerNet BAS datasets. As shown, the heatmaps smoothly follow the ball trajectory across consecutive frames, dynamically shifting their spatial center in accordance with the ball’s movement while maintaining strong temporal localization around the annotated event frame. This motion-aware representation enables the model to precisely highlight the transient interactions that define events such as a volleyball spike or a soccer pass. The results confirm that our formulation successfully encodes both the temporal proximity and the spatial accuracy of discrete events, demonstrating its capability to capture fine-grained spatiotemporal dynamics in complex gameplay scenarios.

5 Conclusion

We introduced the task of Spatiotemporal Event Spotting (STES), which aims to detect discrete events at specific frames along with their corresponding spatial locations in sports videos. Unlike prior methods that focus solely on temporal spotting or action localization over extended durations, STES targets precise, frame-level events and their pixel-level positions, enabling more fine-grained gameplay analysis. To support this task, we presented the VNL dataset, which features spatiotemporal annotations of key volleyball events. Furthermore, we proposed the 3D heatmap representation with dynamically shifted Gaussian kernels, which works effectively with a simple 3D U-Net-based model. Despite its architectural simplicity, our approach achieved state-of-the-art performance on both datasets. We believe that this work provides a strong foundation for future research in spatiotemporal understanding of sports videos and opens new directions for tactical analysis, automated broadcasting, and highlight generation, while also motivating future efforts to improve performance on rare or underrepresented events.

Acknowledgements

This work was supported by Development of foot diagnosis and shoe manufacturing technology for the disabled Project (RS-2022-00164554).

References

- [1] Anurag Arnab, Xuehan Xiong, Alexey Gritsenko, Rob Romijnders, Josip Djolonga, Mostafa Dehghani, Chen Sun, Mario Lučić, and Cordelia Schmid. Beyond transfer learning: Co-finetuning for action localisation. *arXiv preprint arXiv:2207.03807*, 2022.
- [2] Vanyi Chao, Hoang Quoc Nguyen, Ankhzaya Jamsrandorj, Yin May Oo, Kyung-Ryoul Mun, Hyowon Park, Sangwon Park, and Jinwook Kim. Tracking the blur: Accurate ball trajectory detection in broadcast sports videos. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*, pages 41–49, 2024.
- [3] Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan Wu, Lan Ma, Jiajun Shen, and Ping Luo. Watch only once: An end-to-end video action detection framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8178–8187, 2021.
- [4] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.
- [5] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Delière, Jan Held, Carlos Hinojosa, Amir M Mansourian, et al. Soccernet 2023 challenges results. *Sports Engineering*, 27(2):24, 2024.
- [6] Adrien Delière, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. Soccernet-v2 : A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021.
- [7] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. 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*, pages 530–540, 2024.
- [8] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 768–784. Springer, 2016.
- [9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.
- [10] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721, 2018.

- [11] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 244–253, 2019.
- [12] Alexey A Gritsenko, Xuehan Xiong, Josip Djolonga, Mostafa Dehghani, Chen Sun, Mario Lucic, Cordelia Schmid, and Anurag Arnab. End-to-end spatio-temporal action localisation with video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18373–18383, 2024.
- [13] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018.
- [14] Fabian Caba Heilbron, Juan Carlos Nieves, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1914–1923, 2016.
- [15] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *European Conference on Computer Vision*, pages 33–51. Springer, 2022.
- [16] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3320–3329, 2021.
- [17] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019.
- [18] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022.
- [19] Yin May Oo, Ankhzaya Jamsrandorj, Vanyi Chao, Kyung-Ryoul Mun, and Jinwook Kim. A residual attention-based efficientnet homography estimation model for sports field registration. In *IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–7. IEEE, 2023.
- [20] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021.
- [21] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 485–494, 2021.

- [22] Joao VB Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2796–2800. IEEE, 2022.
- [23] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 71–87. Springer, 2020.
- [24] Adrian Wolny, Lorenzo Cerrone, Athul Vijayan, Rachele Tofanelli, Amaya Vilches Barro, Marion Louveaux, Christian Wenzl, Sören Strauss, David Wilson-Sánchez, Rena Lymbouridou, et al. Accurate and versatile 3d segmentation of plant tissues at cellular resolution. *Elife*, 9:e57613, 2020.
- [25] Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. Astra: An action spotting transformer for soccer videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 93–102, 2023.
- [26] Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. 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*, pages 3410–3419, 2024.
- [27] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10156–10165, 2020.
- [28] Le Yang, Junwei Han, Tao Zhao, Nian Liu, and Dingwen Zhang. Structured attention composition for temporal action localization. *IEEE Transactions on Image Processing*, 2022.
- [29] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13598–13607, 2022.