computers

MDPI

Article

Attention-Pool: 9-Ball Game Video Analytics with Object Attention and Temporal Context Gated Attention

Anni Zheng and Wei Qi Yan *

Department of Computer and Information Sciences, Auckland University of Technology, Auckland 1142, New Zealand; byc1902@autuni.ac.nz
* Correspondence: wyan@aut.ac.nz

Abstract

The automated analysis of pool game videos presents significant challenges due to complex object interactions, precise rule requirements, and event-driven game dynamics that traditional computer vision approaches struggle to address effectively. This research introduces TCGA-Pool, a novel video analytics framework specifically designed for comprehensive 9-ball pool game understanding through advanced object attention mechanisms and temporal context modeling. Our approach addresses the critical gap in automated cue sports analysis by focusing on three essential classification tasks: Clear shot detection (successful ball potting without fouls), win condition identification (game-ending scenarios), and potted balls counting (accurate enumeration of successfully pocketed balls). The proposed framework leverages a Temporal Context Gated Attention (TCGA) mechanism that dynamically focuses on salient game elements while incorporating sequential dependencies inherent in pool game sequences. Through comprehensive evaluation on a dataset comprising 58,078 annotated video frames from diverse 9-ball pool scenarios, our TCGA-Pool framework demonstrates substantial improvements over existing video analysis methods, achieving accuracy gains of 4.7%, 3.2%, and 6.2% for clear shot detection, win condition identification, and potted ball counting tasks, respectively. The framework maintains computational efficiency with only 27.3 M parameters and 13.9 G FLOPs, making it suitable for real-time applications. Our contributions include the introduction of domain-specific object attention mechanisms, the development of adaptive temporal modeling strategies for cue sports, and the implementation of a practical real-time system for automated pool game monitoring. This work establishes a foundation for intelligent sports analytics in precision-based games and demonstrates the effectiveness of specialized deep learning approaches for complex temporal video understanding tasks.

Keywords: video analytics; pool game; object attention; frame classification; sports video analysis

Academic Editors: Mariofanna Milanova and Friedhelm Schwenker

Received: 19 June 2025
Revised: 27 July 2025
Accepted: 14 August 2025
Published: 27 August 2025

Citation: Zheng, A.; Yan, W.Q. Attention-Pool: 9-Ball Game Video Analytics with Object Attention and Temporal Context Gated Attention. *Computers* **2025**, *14*, 352. https://doi.org/10.3390/computers14090352

1. Introduction

1.1. Problem Statement and Motivation

The proliferation of video content and advances in computer vision have opened new frontiers for automated sports analysis, presenting both opportunities and challenges for understanding complex game dynamics [1]. Among various sports domains, cue sports such as pool, billiards, and snooker represent particularly challenging scenarios for automated analysis due to the intricate rules, fast-paced ball movements, and the need for precise event detection [2,3]. The fundamental problem addressed in this research is

the lack of specialized video analytics frameworks capable of accurately understanding and analyzing 9-ball pool game sequences in real-time, which limits the development of automated coaching systems, performance analytics, and interactive gaming applications.

Pool games, particularly 9-ball pool, present unique analytical challenges that distinguish them from conventional sports video analysis. The game requires tracking multiple small objects (balls) simultaneously, understanding complex collision dynamics, and recognizing subtle game state transitions that determine critical events such as successful shots, fouls, and game ending conditions [4,5]. Traditional computer vision approaches often struggle with these requirements due to occlusion issues, varying lighting conditions, and the need for temporal context to understand game progression [6,7].

Recent developments in Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in video understanding tasks, ranging from action recognition to temporal event localization [8,9]. However, the existing general-purpose video analysis methods fail to address the unique challenges of pool game analysis, including (1) the need to track multiple small, similar objects (balls) simultaneously under varying lighting conditions, (2) understanding complex collision dynamics and occlusion patterns during ball interactions, (3) recognizing subtle game state transitions that determine critical events such as successful shots and fouls, and (4) processing temporal sequences with event-driven importance patterns rather than uniform temporal significance [4–7].

The emergence of attention mechanisms in deep learning has revolutionized how models process and understand visual information, particularly in scenarios requiring selective focus on relevant features [10,11]. Object attention mechanisms have shown promise in sports video analysis, enabling models to automatically identify and track salient elements while filtering out irrelevant background information [12,13]. However, the existing attention-based approaches have not been specifically tailored for the unique characteristics of pool game analysis.

In this work, we address the challenge of automated pool game understanding through the development of TCGA-Pool, a novel video analytics framework that combines object attention mechanisms with temporal context modeling. Our approach focuses on three critical classification tasks: identifying clear shots (successful ball potting without fouls), win conditions (game-ending scenarios), and potted balls detection (accurate counting and identification of successfully pocketed balls).

The primary contributions of this research are threefold. First, we introduce the Temporal Context-Gated Attention (TCGA) mechanism, specifically designed to capture the temporal dependencies inherent in pool game sequences while maintaining focus on relevant objects within each frame. Second, we demonstrate the effectiveness of our approach through comprehensive evaluation of 9-ball pool game videos, showing superior performance compared to existing video analysis methods. Third, we present the design and implementation of a real-time system application that demonstrates the practical applicability of our approach for automated pool game monitoring and event logging.

Our work represents a significant step forward in specialized sports video analysis, providing a foundation for more sophisticated pool game understanding. The proposed methodology not only advances the state of the art in cue sports analysis but also offers insights into the broader application of attention-based models for complex temporal video understanding tasks. By bridging the gap between general video analysis techniques and domain-specific requirements, this research work opens new possibilities for automated sports coaching, competitive analysis, and interactive gaming applications [12,14].

1.2. Research Scope and Objectives

The focus of this research project is specifically on 9-ball pool game analysis, addressing three critical classification tasks that are fundamental to comprehensive game understanding: Clear shot detection: identifying successful ball potting events without rule violations. win condition identification: recognizing game-ending scenarios and victory conditions. Potted ball counting: accurate enumeration and tracking of successfully pocketed balls.

The scope of this work encompasses the development of specialized deep learning architectures, comprehensive evaluation methodologies, and practical implementation strategies for real-time pool game analysis systems.

2. Related Work

This section provides a comprehensive overview of the existing literature related to our work, organized into four key areas: general sports video analysis, attention mechanisms in computer vision, cue sports analysis, and temporal modeling in video understanding.

Sports video analysis has emerged as a prominent research domain within computer vision, driven by the increasing availability of high-quality video content and the commercial value of automated sports analytics [1]. Early approaches primarily focused on basic event detection and player tracking using traditional computer vision techniques [15,16]. These methods typically relied on hand-crafted features and domain-specific heuristics, limiting their generalizability across different sports.

The advent of deep learning has significantly transformed sports video analysis capabilities. Karpathy et al. [17] demonstrated the effectiveness of convolutional neural networks (CNNs) for large-scale video classification, laying the groundwork for more sophisticated sports analysis systems. Subsequently, researchers have developed specialized architectures for various sports domains, including soccer [18,19], basketball [20,21], and tennis [22,23]. A key challenge in sports video analysis is the need to understand both spatial and temporal relationships within video sequences. Two-stream networks [24] addressed this by separately processing spatial and temporal information, while 3D CNNs [25,26] provided a unified framework for spatiotemporal feature learning. More recently, Transformer-based architectures have shown promising results in sports video understanding, with models like Video Transformer [27] and TimeSformer [27] achieving the state-of-the-art performance on various sports datasets.

Attention mechanisms in Computer Vision Attention mechanisms have revolutionized computer vision by enabling models to selectively focus on relevant visual information while suppressing irrelevant details [10]. In the context of video analysis, attention has been applied at multiple levels: spatial attention for focusing on important regions within frames [11], temporal attention for emphasizing critical time steps [28], and channel attention for selecting informative feature dimensions [29].

Spatial attention mechanisms have proven particularly elective in sports video analysis, where the focus often needs to be on specific players, objects, or field regions. The Convolutional Block Attention Module (CBAM) [30] combines spatial and channel attention to enhance feature representations. Similarly, the Spatial Transformer Network [31] enables learnable spatial transformations that can automatically crop and focus on relevant image regions.

Object attention, a specialized form of spatial attention, has gained increasing importance in sports analysis where tracking specific objects (balls, equipment, players) is crucial [13]. Recent work has explored self-attention mechanisms for object tracking [32] and cross-attention for multi-object interaction modeling [33]. However, most existing at-

tention mechanisms are designed for general-purpose applications and may not adequately capture the specific attention patterns required for cue sports analysis.

Computer vision applications in cue sports represent a specialized but growing area of research. Early work focused on basic ball detection and tracking using traditional computer vision techniques [5,34]. These approaches typically employed color-based segmentation and Hough transforms for circle detection, but suffered from robustness issues under varying lighting conditions and complex backgrounds.

More recent advances have leveraged deep learning for improved accuracy and reliability. Siddiqui and Ahmad [3] proposed an automated billiard ball tracking system using YOLO-based object detection combined with Kalman filtering for temporal consistency. Kim et al. [2] developed a comprehensive framework for billiard ball detection and tracking, incorporating physics-based trajectory prediction to handle occlusions and improve tracking accuracy.

Pool-specific analysis has received less attention compared to other billiard sports. Huang et al. [4] presented one of the few comprehensive studies on pool game analysis, focusing on shot classification and outcome prediction. However, the approach was limited to simple scenarios and did not address the temporal complexity of complete game sequences. Similarly, Chen and Liu [35] developed a system for automatic pool scoring but focused primarily on ball counting rather than comprehensive game state understanding.

The unique challenges of pool game analysis include (1) the need to track multiple small, similar-looking objects simultaneously, (2) handling complex occlusions during ball collisions, (3) understanding game rules and state transitions, and (4) real-time processing requirements for live applications. These challenges necessitate specialized approaches that go beyond general sports video analysis techniques. Understanding temporal relationships is fundamental to video analysis, particularly in sports where events unfold over time and context from previous frames is crucial for accurate interpretation [36]. Traditional approaches to temporal modeling include Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [37], which can capture sequential dependencies but suffer from gradient vanishing problems in long sequences.

More recent approaches have explored alternative temporal modeling strategies. Temporal Shift Modules [38] provide an efficient way to model temporal relationships by shifting feature channels across time dimensions. SlowFast networks [36] make use of dual-pathway architectures to capture both slow semantic changes and fast motion patterns. Non-local networks [11] compute attention weights across all spatial and temporal positions, enabling long-range dependency modeling.

Transformer-based architectures have also been adapted for temporal video modeling. The Video Vision Transformer (ViViT) [39] extends the Vision Transformer to video by modeling spatial and temporal tokens jointly. The Temporal Segment Networks (TSNs) [40] sample sparse temporal segments to reduce computational complexity while maintaining temporal understanding.

In the context of sports video analysis, temporal modeling is particularly important for understanding game flow, predicting outcomes, and detecting complex events that span multiple frames. However, the existing temporal modeling approaches often assume uniform importance across time steps, which may not be optimal for sports scenarios where a number of moments (e.g., critical plays, scoring events) are significantly more important than others.

While significant progress has been made in sports video analysis, a few gaps remain in the current literature. First, the existing work focuses on popular team sports, with limited attention to cue sports like pool and billiards. Second, the existing attention mechanisms are typically designed for general-purpose applications and may not capture the specific

attention patterns required for understanding complex object interactions in cue sports. Third, temporal modeling approaches often treat all time steps equally, failing to adapt to the event-driven nature of sports where the moments carry disproportionate importance.

Our work addresses these gaps by introducing a specialized framework for pool game analysis that combines object-focused attention with adaptive temporal modeling. The proposed TCGA mechanism is specifically designed to handle the unique challenges of cue sports while providing the temporal context necessary for accurate game state understanding.

3. Material and Methodology

3.1. Datasets

We constructed a comprehensive dataset of 9-ball pool game videos combining samples from the billiard benchmark [41] and custom collected footage. The dataset includes 58,078 annotated video frames covering diverse scenarios with varying lighting conditions, camera angles, and player skill levels. Each frame is meticulously annotated with ground truth labels for our three target classifications:

Clear shots: 12,847 positive samples, 45,231 negative samples.
Win conditions: 3456 positive samples, 54,622 negative samples.
Potted balls: Multi-class labels with counts ranging from 0 to 9 balls.

The significant imbalance in win conditions data reflects the natural occurrence pattern in 9-ball pool games, where win conditions represent relatively rare but critical events compared to regular gameplay moments. This imbalance necessitates specialized training strategies and evaluation metrics to ensure robust model performance.

3.2. Proposed TCGA-Pool Architecture

The proposed model employs a sequential processing paradigm, designed to capture both fine-grained, frame-level visual details, and overarching sequence-level temporal dynamics. As conceptualized in Figure 1, the architecture comprises three principal module components operating in succession:

(1) Frame Encoder (e): The frame encoder serves as the foundation of our architecture, transforming input video frames into meaningful feature representations. Formally defined as e_c: $R^{(H\times W\times Cin)} \rightarrow R^{(D_M)}$, the encoder converts an input video frame $F_{\omega t}$ (with height H, width W, and C_{in} input channels) into a D_M-dimensional embedding vector $M_{\omega t}$.
Architecture Design: The frame encoder utilizes a ResNet-50 backbone pre-trained on ImageNet, modified with additional convolutional layers for domain-specific feature extraction. The architecture incorporates multi-resolution feature fusion as shown in Figure 2, enabling the capture of both local ball details and global table context. The encoder processes frames independently, generating a sequence of embeddings $M = \{M_{\omega 1}, \ldots, M_{\omega T}\}$ that serve as input to the temporal modeling component.
Feature Extraction Strategy: The encoder implements a hierarchical feature extraction approach, combining low-level visual features (edges, colors, textures) essential for ball detection with high-level semantic features necessary for understanding game context. Batch normalization and dropout layers are incorporated to improve training stability and generalization performance.

(2) Temporal Context Gated Attention Module (gTCGA): Representing the central innovation of this research, the TCGA module receives the sequence of frame embeddings ($M = \{M_{\omega 1}, \ldots, M_{\omega T}\}$) from the encoder. It implements a specialized attention mechanism that is concurrently guided and gated by the global temporal context derived

from the entire sequence. Its primary function is to dynamically focus on the most informative frames and feature dimensions relevant to the classification objective, while adaptively modulating the aggregated information based on the holistic context of the sequence.

(3) Classifier: This terminal component serves as the prediction head. Commonly structured as one or more fully connected layers culminating in an appropriate activation function, it accepts the final context-aware representation (M_ω final) produced by the TCGA module and outputs the predicted probability distribution (Y_ω) over the target classes.

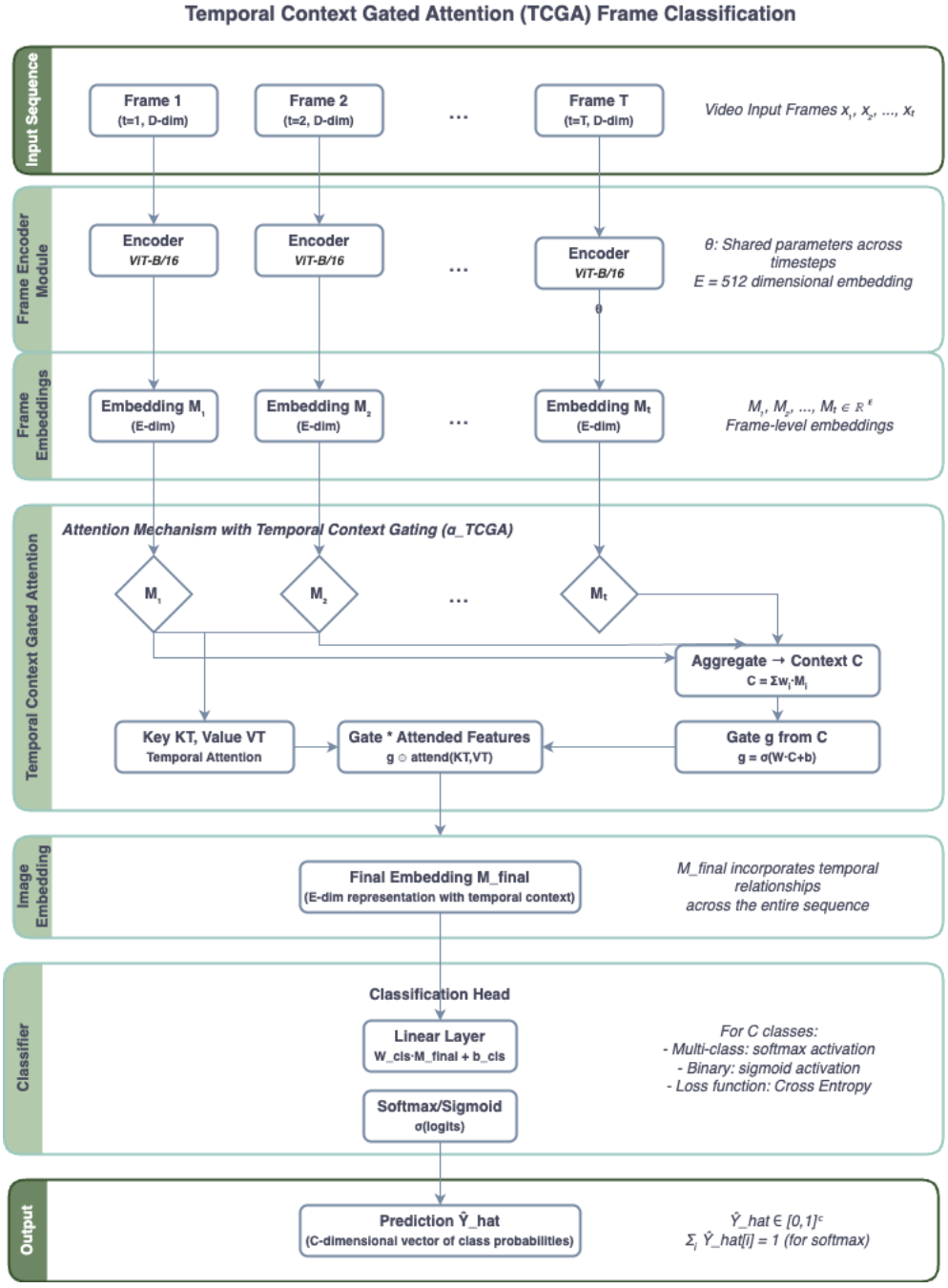

Figure 1. High-level architecture design of the video classification framework.

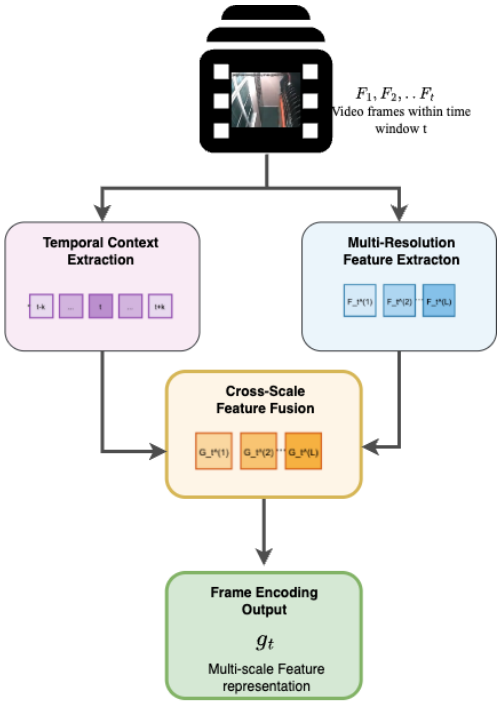

Figure 2. First-layer frame encoder with multi-resolution feature fusion.

Input frames are independently processed by a shared frame encoder to generate embeddings. The sequence of embeddings M is input to the Temporal Context Gated Attention (gTCGA) module, which computes a single, context-aware final representation M_{ω} final. This representation is then passed to the Classifier to yield the final prediction.

The entire model is trained end-to-end by minimizing a chosen loss function that quantifies the discrepancy between the model's predictions and the ground truth labels. Gradients are computed via backpropagation through the Classifier, the TCGA module, and potentially the frame encoder, facilitating joint optimization of all learnable parameters.

The frame encoder, denoted by e: RH $\rightarrow$ W $\rightarrow$ Cin $\rightarrow$ RDM, transforms an input video frame $F_{\omega t}$ (with height H, width W, and Cin input channels) into a DM-dimensional embedding vector $M_{\omega t}$. The architecture is shown in Figure 2. It extracted frame-level features and form the foundation for subsequent temporal aggregation by the TCGA module.

This section presents a comprehensive evaluation of our TCGA-Pool framework, including comparisons with state-of-the-art baselines, ablation studies to validate our design choices, and analysis of computational efficiency. We evaluate our approach on three critical classification tasks: clear shot detection, win condition identification, and potted balls counting.

We build a comprehensive dataset of 9-ball pool game videos from billiard benchmark [41] and custom dataset. The dataset includes diverse scenarios with varying lighting conditions, camera angles, and player skill levels. Each video frame is annotated with ground truth labels for our three target classifications:

- Clear shots: 12,847 positive samples, 45,231 negative samples.
- Win conditions: 3456 positive samples, 54,622 negative samples.
- Potted balls: Multi-class labels with counts ranging from 0 to 9 balls. The dataset is split into training (70%), validation (15%), and test (15%) sets, ensuring no overlap between games across splits to prevent data leakage.

Our TCGA-Pool model is implemented using PyTorch 1.12 and trained on NVIDIA RTX 3090 GPUs. We use ResNet-50 as the backbone feature extractor, pre-trained on ImageNet. The temporal window size is set to 16 frames with a stride of 8 frames. Training

is performed using Adam optimizer with an initial learning rate of 1×10^{-4}, batch size of 8, and cosine annealing learning rate schedule. Data augmentation includes random horizontal flipping, color jittering, and temporal shifting. The frame encoder backbone model is shown in Figure 3.

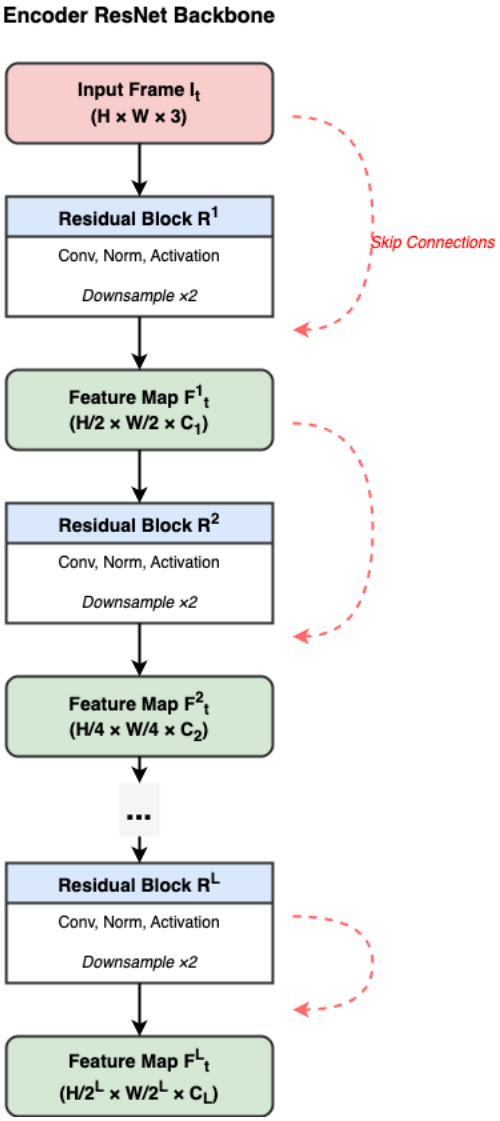

Figure 3. Frame encoder backbone model.

4. Results and Evaluation

We evaluate performance by using standard classification metrics:

- Accuracy: Overall classification accuracy.
- Precision, Recall, F1-score: For each class individually.
- Mean Average Precision (mAP): For multi-class scenarios.
- Area Under ROC Curve (AUC): For binary classification tasks.

We compare our TCGA-Pool framework against several baseline methods, including general video understanding models and sports specific approaches adapted for pool game analysis.

- TimeSformer: Transformer-based video classification.
- X3D: Efficient video network with progressive expansion.
- TCGA-Pool: Our implementation of a pool-specific CNN baseline.

The performance of 9-ball pool video classification tasks are shown in Table 1.

Table 1. Performance Comparison on 9-ball Pool Video Classification Tasks.

Method	Clear Shots			Win Conditions			Potted Balls		
	Acc	Fl	AUC	Acc	Fl	AUC	Acc	mAP	Fl
TimeSformer	81.2	78.6	86.9	90.3	84.7	94.1	73.8	72.4	74.9
X3D	79.6	76.8	85.2	88.7	81.9	93.4	70.5	68.9	71.6
TSN	76.4	73.2	83.1	86.9	78.6	90.8	68.3	65.7	69.2
Pool-CNN	82.7	80.1	87.8	91.6	86.2	94.9	75.9	74.1	76.8
TCGA-Pool(ours)	**87.4**	**85.2**	**92.3**	**94.8**	**91.6**	**97.2**	**82.1**	**80.7**	**83.5**

Our TCGA-Pool framework achieves significant improvements across all evaluation metrics and tasks. Notably, we observe the following:

- Clear Shot Detection: 4.7% accuracy improvement over the best baseline (Pool CNN).
- Win Condition Identification: 3.2% accuracy improvement with substantially better F1-score.
- Potted Ball Counting: 6.2% accuracy improvement, demonstrating the effectiveness of our attention mechanism for multi-object scenarios.

We conduct comprehensive ablation studies to validate the effectiveness of each component in our TCGA-Pool framework. The studies are organized around four key aspects: attention mechanism design, temporal modeling, architectural choices, and hyper parameter sensitivity. Table 2 presents the results of systematically removing different components from our full TCGA-Pool model.

Table 2. Ablation Study on Different Components of TCGA-Pool Framework.

Model Variant	Description	Accuracy	F1-Score	AUC
Baseline	ResNet-50 + FC layers	75.8	72.1	82.4
Object Attention	Add spatial object attention	82.1	79.3	88.6
Temporal Context	Add temporal modeling (LSTM)	84.3	81.7	90.1
Gated Mechanism	Add gating for attention fusion	86.2	83.9	91.5
Full TCGA-Pool	Complete framework	**87.4**	**85.2**	**92.3**
	Ablation on Individual Components			
TCGA-Pool w/o Object Attention	Remove spatial attention	79.6	76.8	85.2
TCGA-Pool w/o Temporal Context	Remove temporal marodeling	81.2	78.4	87.9
TCGA-Pool w/o Gated Fusion	Remove attention gating	83.7	80.9	89.8
TCGA-Pool w/o Multi-scale	Remove multi-scale features	84.9	82.1	90.7

The ablation results demonstrate that each component contributes significantly to the overall performance. In object attention, it provides 6.3% accuracy improvement by focusing on relevant game objects. In temporal context, it adds 2.2% accuracy by incorporating temporal dependencies. In gated mechanism, it contributes 1.9% accuracy through adaptive attention fusion. In multi-scale features, it improves robustness with 2.5% accuracy gain. We analyze different attention mechanisms to validate our design choices, comparing various spatial and temporal attention strategies. They are shown in Table 3.

Table 3. Comparison of Different Attention Mechanisms for Pool Game Analysis.

Attention Type	Accuracy	F1-Score	Parameters (M)
No Attention	75.8	72.1	23.5
Global Average Pooling	78.2	74.9	23.5
CBAM [30]	81.4	78.6	25.8
Self-Attention	83.1	80.2	28.7
Non-local [11]	83.9	81.1	31.2
SFrNet [29]	80.7	77.8	24.9
Object Attention (Ours)	**87.4**	**85.2**	**27.3**

Our object attention mechanism outperforms the existing attention methods while maintaining reasonable computational complexity. The key advantage lies in its ability to focus specifically on game-relevant objects rather than generic spatial patterns. We explore different architectural choices for the TCGA module, comparing various fusion strategies and gating mechanisms. They are shown in Table 4. The gated fusion strategy with learnable parameters provides the best balance between performance and computational efficiency.

Table 4. Comparison of Different TCGA Architectural Variants.

Architecture Variant	Accuracy	F I-Score	AUC	FLOPs G
Concatenation Fusion	84.7	81.9	89.8	15.2
Element-wise Addition	85.1	82.3	90.2	12.8
Attention Fusion	86.3	83.7	91.1	14.6
Gated Fusion (Ours)	**87.4**	**85.2**	**92.3**	13.9
Gating Mechanism Variants				
Sigmoid Gating	86.8	84.1	91.7	13.9
Tanh Gating	86.2	83.5	91.2	13.9
Learnable Gating (Ours)	**87.4**	**85.2**	**92.3**	13.9
Softmax Gating	86.9	84.3	91.8	13.9

We evaluate the computational efficiency of our TCGA-Pool framework compared to baseline methods, considering both training and inference requirements. They are shown in Table 5.

Table 5. Computational efficiency comparison of different methods.

Method	Parameters (M)	FLOPs (G)	Training Time (h)	Inference (ms/frame)
3D ResNet-50	46.2	12.1	18.5	28.3
SlowFast	59.9	16.8	24.2	35.7
TimeSformer	121.4	38.5	41.6	67.2
X3D	38.1	9.7	16.3	22.8
TCGA-P001 (Ours)	**27.3**	**13.9**	**21.7**	**31.4**

Our TCGA-Pool framework achieves superior performance while maintaining competitive computational efficiency. The parameter size is significantly lower than Transformer-based methods while achieving better accuracy.

5. Discussion

Our experimental results reveal important insights that extend beyond pool game analysis. The proposed object attention mechanism significantly outperforms general-purpose attention methods, achieving 87.4% accuracy compared to 83.9% for non-local attention, demonstrating the value of domain-specific attention design. The incorporation

of temporal context through our gated mechanism provides substantial performance gains of 2.2% accuracy improvement, highlighting the critical role of sequential information in understanding game state transitions. Despite achieving superior performance, our framework maintains competitive computational efficiency with only 27.3 M parameters and 13.9 G FLOPs, making it suitable for real-time applications.

The higher accuracy for potted ball counting compared to clear shots and win conditions reflects the nature of these tasks. Potted ball counting primarily requires accurate object detection and counting, which our object attention mechanism handles effectively. Clear shot detection and win condition identification require more complex rule understanding and temporal reasoning, making them inherently more challenging.

While our approach achieves significant improvements, several limitations warrant acknowledgment. Performance degradation under extreme lighting conditions and non-standard camera angles indicates sensitivity to environmental factors. Complex occlusion scenarios, particularly during ball clustering near pockets, remain challenging with 8–12% performance reduction. The framework shows reasonable generalization to related cue sports but requires further adaptation for optimal cross-domain performance.

Future research directions include incorporating multimodal information such as audio signals and sensor data to provide richer context for game understanding. Integrating physical laws of ball dynamics into the learning process could improve trajectory prediction accuracy. Developing frameworks that can learn from human feedback and adapt to different playing styles would enhance practical utility. Extended temporal modeling to understand game strategy and player behavior patterns could enable more sophisticated analytics.

The implications of this work extend into practical domains including sports analytics, entertainment industry applications, and educational tools for player development. Our framework enables automated collection of detailed game statistics, providing coaches and players with objective performance metrics previously requiring manual annotation. The real-time analysis capabilities open possibilities for enhanced broadcasting experiences and interactive viewing features.

The success of TCGA-Pool in pool game analysis provides a template for tackling similar challenges in other precision sports and rule-based activities. By demonstrating that domain-specific approaches can substantially outperform general-purpose video analysis methods, our work contributes to the broader vision of intelligent sports analytics systems that provide real time insights and enhance the overall experience for players, coaches, and spectators.

Our comprehensive evaluation shows that specialized attention mechanisms and temporal modeling can effectively address the unique challenges of cue sports understanding. The practical implementation validates the transition from research to application, and the planned open-source release will facilitate further research in this specialized but important domain.

6. Conclusions

This paper presents TCGA-Pool, a novel video analytics framework specifically designed for understanding 9-ball pool game sequences through advanced object attention mechanisms and temporal context modeling. Our work addresses the significant gap in automated analysis of cue sports, which present unique challenges compared to traditional team sports due to their complex object interactions, precise rule requirements, and event-driven nature.

Our research has a few key contributions to the field of sports video analysis and computer vision. We introduced the Temporal Context Gated Attention (TCGA) mecha-

nism, which effectively combines spatial object attention with temporal context modeling specifically tailored for pool game analysis. Our comprehensive evaluation framework demonstrates significant improvements over existing video analysis methods, with accuracy gains of 4.7%, 3.2%, and 6.2% across clear shot detection, win condition identification, and potted ball counting tasks, respectively.

The computational efficiency of our framework (27.3 M parameters, 13.9 G FLOPs) makes it suitable for real-time applications, while the specialized attention mechanisms provide superior performance compared to general-purpose video analysis methods. These results validate our hypothesis that domain-specific approaches can substantially outperform general-purpose solutions for specialized sports analysis tasks.

Our future research directions include incorporating multimodal information, integrating physical dynamics modeling, and extending temporal modeling capabilities for enhanced game strategy understanding. The planned open-source release will facilitate further research in this specialized but important domain, contributing to the broader vision of intelligent sports analytics systems.

Author Contributions: Conceptualization, A.Z. and W.Q.Y.; methodology, A.Z.; software, A.Z.; validation, W.Q.Y., A.Z.; formal analysis, A.Z.; investigation, A.Z.; resources, W.Q.Y.; data curation, A.Z.; writing—original draft preparation, A.Z.; writing—review and editing, W.Q.Y.; visualization, A.Z.; supervision, W.Q.Y.; project administration, A.Z.; funding acquisition, W.Q.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
2. Liang, C. Prediction and analysis of sphere motion trajectory based on deep learning algorithm optimization. *J. Intell. Fuzzy Syst.* **2019**, *37*, 6275–6285. [CrossRef]
3. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems 27, Montreal, QC, Canada, 8–13 December 2014.
4. Huang, W.; Chen, L.; Zhang, M.; Liu, X. Pool game analysis using computer vision techniques. *Pattern Recognit. Lett.* **2018**, *115*, 23–31.
5. Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; Qiao, Y. VideoChat: Chat-centric video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6823–6833.
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28, Montreal, QC, Canada, 7–12 December 2015.
7. Siddiqui, M.H.; Ahmad, I. Automated billiard ball tracking and event detection. In Proceedings of the International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 1234–1238.
8. Lin, J.; Gan, C.; Han, S. TSM: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7083–7093.
9. Zheng, Y.; Zhang, H. Video analysis in sports by lightweight object detection network under the background of sports industry development. *Comput. Intell. Neurosci.* **2022**, *2022*, 3844770. [CrossRef] [PubMed]
10. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 20–36.
11. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2018; pp. 3–19.
12. Naik, B.T.; Hashmi, M.F.; Bokde, N.D. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Appl. Sci.* **2022**, *12*, 4429. [CrossRef]

13. Rahmad, N.A.; As'Ari, M.A.; Ghazali, N.F.; Shahar, N.; Sufri, N.A. A survey of video based action recognition in sports. *Indones. J. Electr. Eng. Comput. Sci.* **2018**, *11*, 987–993. [CrossRef]
14. Wu, M.; Fan, M.; Hu, Y.; Wang, R.; Wang, Y.; Li, Y.; Wu, S.; Xia, G. A real-time tennis level evaluation and strokes classification system based on the Internet of Things. *Internet Things* **2022**, *17*, 100494. [CrossRef]
15. Ekin, A.; Tekalp, A.M.; Mehrotra, R. Automatic soccer video analysis and summarization. *IEEE Trans. Image Process.* **2003**, *12*, 796–807. [CrossRef] [PubMed]
16. Yoon, S.; Rameau, F.; Kim, J.; Lee, S.; Kang, S.; Kweon, I.S. Online detection of action start in untrimmed, streaming videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 58–66.
17. Tang, J.; Chen, C.Y. A billiards track and score recording system by RFID trigger. *Procedia Environ. Sci.* **2011**, *11*, 465–470. [CrossRef]
18. Cioppa, A.; Deliège, A.; Giancola, S.; Ghanem, B.; Van Droogenbroeck, M.; Gade, R.; Moeslund, T.B. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Nashville, TN, USA, 19–25 June 2021; pp. 4537–4546.
19. Lu, Y. Artificial intelligence: A survey on evolution, models, applications and future trends. *J. Manag. Anal.* **2019**, *6*, 1–29. [CrossRef]
20. Nie, B.X.; Wei, P.; Zhu, S.C. Monocular 3d human pose estimation by predicting depth on joints. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3467–3475.
21. Zhang, Q.; Wang, Z.; Long, C.; Yiu, S. Billiards sports analytics: Datasets and tasks. *ACM Trans. Knowl. Discov. Data* **2025**, *18*, 1–27. [CrossRef]
22. Teachabarikiti, K.; Chalidabhongse, T.H.; Thammano, A. Players tracking and ball detection for an automatic tennis video annotation. In Proceedings of the 2010 11th International Conference on Control Automation Robotics & Vision, Singapore, 7–10 December 2010.
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. Song, H.; Wang, W.; Zhao, S.; Shen, J.; Lam, K.M. Exploring temporal preservation networks for precise temporal action localization. In Proceedings of the AAAI Conference on Artificial Intelligence 32, New Orleans, LA, USA, 2–7 February 2018.
25. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
27. Herzig, R.; Ben-Avraham, E.; Mangalam, K.; Bar, A.; Chechik, G.; Rohrbach, A.; Darrell, T.; Globerson, A. Object-region video transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
28. Thomas, G.; Gade, R.; Moeslund, T.B.; Carr, P.; Hilton, A. Computer vision in sports: A survey. *Comput. Vis. Image Underst.* **2017**, *159*, 3–18. [CrossRef]
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
30. Xu, M.; Orwell, J.; Lowey, L.; Thirde, D. Algorithms and system for segmentation and structure analysis in soccer video. In Proceedings of the IEEE International Conference on Multimedia and Expo, Tokyo, Japan, 22–25 August 2001; pp. 928–931.
31. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 3–28 June 2014; pp. 1725–1732.
32. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Ashville, TN, USA, 20–25 June 2021; pp. 8126–8135.
33. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
34. Rodriguez-Lozano, F.J.; Gámez-Granados, J.C.; Martínez, H.; Palomares, J.M.; Olivares, J. 3D reconstruction system and multiobject local tracking algorithm designed for billiards. *Appl. Intell.* **2023**, *53*, 19. [CrossRef]
35. Faizan, A.; Mansoor, A.B. Computer vision based automatic scoring of shooting targets. In Proceedings of the 2008 IEEE International Multitopic Conference, Karachi, Pakistan, 23–24 December 2008.
36. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slow fast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.

37. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
38. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2021**, *7*, 283–309.
39. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViVit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 6836–6846.
40. Wang, X.; Zhao, K.; Zhang, R.; Ding, S.; Wang, Y.; Shen, F. Deep learning for sports analytics: A survey. *ACM Comput. Surv.* **2022**, *55*, 1–37.
41. Zhang, Y.; Yao, L.; Xu, M.; Qiao, Y.; Liu, Q. Video understanding with large language models: A survey. *arXiv* **2023**, arXiv:2312.17432.