applied sciences

MDPI

Article

Sports Video Classification Method Based on Improved Deep Learning

Tianhao Gao [1,†], Meng Zhang [1,†], Yifan Zhu [2], Youjian Zhang [1], Xiangsheng Pang [1], Jing Ying [2] and Wenming Liu [1,*]

[1] Department of Sport Science, College of Education, Zhejiang University, Hangzhou 310027, China; 22303061@zju.edu.cn (T.G.); 0923473@zju.edu.cn (M.Z.); zyj15416@gmail.com (Y.Z.); pangxs@zju.edu.cn (X.P.)
[2] College of Computer Science, Zhejiang University, Hangzhou 310027, China; xtf_z@zju.edu.cn (Y.Z.); csyingj@zju.edu.cn (J.Y.)
* Correspondence: liuwenming@zju.edu.cn
† These authors contributed equally to this work.

Abstract: Classifying sports videos is complex due to their dynamic nature. Traditional methods, like optical flow and the Histogram of Oriented Gradient (HOG), are limited by their need for expertise and lack of universality. Deep learning, particularly Convolutional Neural Networks (CNNs), offers more effective feature recognition in sports videos, but standard CNNs struggle with fast-paced or low-resolution sports videos. Our novel neural network model addresses these challenges. It begins by selecting important frames from sports footage and applying a fuzzy noise reduction algorithm to enhance video quality. The model then uses a bifurcated neural network to extract detailed features, leading to a densely connected neural network with a specific activation function for categorizing videos. We tested our model on a High-Definition Sports Video Dataset covering over 20 sports and a low-resolution dataset. Our model outperformed established classifiers like DenseNet, VggNet, Inception v3, and ResNet-50. It achieved high precision (0.9718), accuracy (0.9804), F-score (0.9761), and recall (0.9723) on the high-resolution dataset, and significantly better precision (0.8725) on the low-resolution dataset. Correspondingly, the highest values on the matrix of four traditional models are: precision (0.9690), accuracy (0.9781), F-score (0.9670), recall (0.9681) on the high-resolution dataset, and precision (0.8627) on the low-resolution dataset. This demonstrates our model's superior performance in sports video classification under various conditions, including rapid motion and low resolution. It marks a significant step forward in sports data analytics and content categorization.

Keywords: sports video analysis; deep learning; Convolutional Neural Networks (CNN); image processing

Citation: Gao, T.; Zhang, M.; Zhu, Y.; Zhang, Y.; Pang, X.; Ying, J.; Liu, W. Sports Video Classification Method Based on Improved Deep Learning. *Appl. Sci.* **2024**, *14*, 948. https://doi.org/10.3390/app14020948

Academic Editor: Andrea Prati

Received: 17 December 2023
Revised: 15 January 2024
Accepted: 19 January 2024
Published: 22 January 2024

1. Introduction

The precise and expedient categorization of sports videos holds pivotal significance across a spectrum of applications [1], encompassing automated sports analytics [2,3], content-centric retrieval [4], event distillation [5], and the orchestration of tailored recommendation systems [6]. Given the meteoric surge in sports video content coupled with the escalating demand for sophisticated analytical paradigms, the performance constraints of previous methodologies [3,7] are becoming palpably discernible. Consequently, there emerges an imperative for avant-garde and resilient strategies to navigate the intricate challenges intrinsic to the multifaceted and kinetic realm of sports endeavors [8,9].

Conventional methodologies for sports video categorization have predominantly hinged upon manually engineered descriptors, encompassing optical flow [10] and the Histogram of Oriented Gradients (HOG) [11], to discern and delineate particular actions and dynamics within sports footage.

These handcrafted features, while historically pivotal, manifest inherent constraints across diverse tasks and domains. This becomes particularly salient within the sports arena, where the juxtaposition of rapid action dynamics and the nuances of low-resolution

video attributes accentuates these limitations. To elucidate: (i) Temporal Intensity: The crafting of such features demands an extensive reservoir of domain expertise and timely investment. This translates to potentially weeks or even months dedicated to the design and fine-tuning of these features for each novel task or domain. For instance, architecting a feature tailored for football match analysis necessitates profound insights into the sport's intricacies and iterative experimentation. (ii) Task Specificity: Given the bespoke nature of these features, tailored for distinct tasks or domains, their efficacy may wane when transposed to alternative contexts. Drawing from the realm of sports, a descriptor sculpted for baseball might falter when applied to basketball or football, given the stark divergence in rules and foundational actions. (iii) Challenges of High-Velocity Movements: Actions that transpire at elevated speeds might span merely a few frames within a video, yet they encapsulate a wealth of information. Manually crafted features might falter in capturing these ephemeral yet pivotal movements, culminating in the omission of crucial data. For instance, in table tennis, a game-altering rapid spin might elude capture through traditional handcraft techniques. (iv) Pitfalls with Low-Resolution Footage: The detail attrition inherent to low-resolution videos presents another formidable challenge for handcrafted features. For instance, when revisiting and analyzing archival footage from seminal events, constrained by the technological limitations of yesteryear recording apparatuses, one grapples with the intricacies of processing and recognizing low-resolution video samples. Herein, handcrafted features often stumble, leading to detail erosion.

Such challenges can culminate in harmful outcomes, like skewed system predictions and resource misallocation. Take football match analysis as an example: if a researcher manually crafts a feature to monitor player movements, the confluence of rapid player motion during pivotal goal opportunities and low video resolution might compromise the feature's ability to accurately pinpoint player positions and actions. This could relay misleading tactical insights to the coaching staff, potentially swaying the match's trajectory. In essence, while manually engineered descriptors might exhibit efficacy under certain circumstances, their inherent limitations and potential pitfalls have catalyzed a shift towards deep learning-based automatic feature extraction techniques, renowned for their robustness amidst intricate and dynamic data.

Consequently, the academic background has witnessed an augmented inclination towards unearthing methodologies that can surmount these impediments. Deep learning, with an emphasis on Convolutional Neural Networks (CNNs), has burgeoned as a formidable contender in the realm of sports video categorization, showcasing exemplary prowess in image classification endeavors. CNNs harbor the potential for fine-tuned adaptation to video datasets, facilitating the discernment of distinct actions and dynamics inherent to diverse sporting disciplines.

Although traditional CNN architectures, such as VGG [12] and ResNet [13], have garnered significant accolades across a plethora of visual tasks, they encounter specific challenges when navigating sports videos replete with high-velocity movements. Firstly, conventional CNNs, having been trained on high-resolution image datasets, might witness performance degradation when confronted with low-resolution or suboptimal quality sports videos. The absence of efficacious preprocessing techniques to mitigate these artifacts further jeopardizes classification accuracy. Moreover, VGG and ResNet excel in generic image classification paradigms, they might falter in discerning the fine-tuned features intrinsic to the rapid and intricate movements characteristic of sports videos. Meanwhile, it is also noteworthy that in actual sports matches, the background is often dynamic, and the movement and speed of athletes are fast. Additionally, the broadcast or recording quality of lower-level events may not be guaranteed. Traditional deep learning-based sports video classification methods are susceptible to these factors, which could threaten their reliability. Developing strategies to mitigate these threats remains an open research area. Furthermore, the generalizability of these methods across different sports and video quality levels is limited. When faced with new domains or data distributions, they often result in significant

performance drops. Techniques such as domain adaptation and data augmentation can offer assistance but come with their own set of challenges.

In the field of sports video classification, recent years have seen the development of various studies and methods, especially those focusing on addressing the challenges of rapid motion and low resolution [14–16], which have provided solutions from different perspectives. However, there is still room for improvement in terms of their generalizability and accuracy.

In a bid to surmount the aforementioned challenges, this paper introduces the following enhancements: (i) Bifurcated Neural Network Architecture: This design, juxtaposed against conventional CNNs, exhibits an augmented capability to extract features across diverse scales, thereby adeptly capturing the nuances of high-velocity movements. (ii) Pivotal Frame Sampling: This strategy ensures the model's focus is honed on the most salient segments of the video, striking an optimal equilibrium between precision and computational efficiency. (iii) Fuzzy Noise Attenuation Algorithm: Tailored to address the intricacies of low-resolution footage, this algorithm proficiently mitigates motion-induced artifacts and noise, bolstering the model's classification accuracy. (iv) Multiscale and Multiperspective Feature Extraction: This approach offers a holistic capture of video intricacies, ensuring rapid movements are accurately discerned, even under the constraints of diminished resolution.

Specifically, the real-world key challenges and the corresponding solutions proffered in this study can be delineated as follows:

Image Quality Degradation due to Motion Artifacts: The high-velocity movements inherent to sports videos often induce motion artifacts. These artifacts can blur the imagery, potentially misleading the feature extraction process and culminating in imprecise categorizations.

To address this, our study employs the Laplacian second-order differential linear operator to discern high-frequency components within the input imagery. A scarcity of high-frequency components suggests potential blurring and motion artifacts. Concurrently, assuming the original image has been displaced by L pixels either vertically or horizontally, a Fourier transform is invoked to decompose and rectify this blurring. The Radon transform is leveraged to compute the frequency offset angle, yielding the blur kernel angle. An inverse Fourier transform is then executed to retrieve a pristine image. The Laplacian operator meticulously identifies frames afflicted by blurring, and the Fourier transform technique rectifies these artifacts, ensuring feature extraction transpires on artifact-free, sharp imagery.

Performance Issues of Traditional Feature Extraction Techniques in Low Resolution: Sports videos, often captured under varied conditions, exhibit inconsistent resolutions, occasionally skewing towards the lower end. While conventional neural network architectures flourish on high-resolution imagery, their efficacy wanes on low-resolution footage. The detail attrition and pixel loss in low-resolution videos impede traditional techniques from accurately capturing pivotal visual features, leading to performance degradation.

In response, we introduce a novel bifurcated neural network architecture capable of gleaning granular local features across diverse scales and fields of view. This ensures that even under diminished resolution, pivotal features are captured for precise classification. By synergistically harnessing multi-scale and multi-view feature extraction, our methodology retains efficiency and precision on low-resolution videos, transcending the constraints of traditional approaches.

The proposed dual-branch neural network efficiently fuses diverse and complementary local features, enabling improved classification performance in challenging scenarios, such as high-speed sports and low-resolution videos. We present a comprehensive evaluation of the method using the self-collected High-Definition Sports Video Dataset (DeepSports-VDS), and sports videos with motion artifacts and a lower resolution as an additional validation dataset.

Our key contributions are as follows: (i) Problem-driven Approach: We introduced a novel approach that specifically addresses the challenges of classifying sports videos, especially when dealing with high-speed movements and low-resolution inputs. (ii) Noise and Motion Artifact Reduction: Leveraging the Laplacian second-order differential linear operator, we effectively identified and removed motion artifacts from sports

videos, thereby enhancing the clarity of images, especially in high-speed sports scenarios. (iii) Multiscale and Multiperspective Feature Extraction: Our unique dual-branch neural network architecture facilitates the extraction of fine-grained local features across multiple scales and fields of view. This ensures the capture of essential features even in low-resolution scenarios, leading to more accurate classifications. (iv) Better Performance across Different Resolutions: On the High-Definition Sports Video Dataset (DeepSports-VDS), our method achieved an unparalleled accuracy of 0.9804, surpassing the ResNet-50, by a substantial 1%. Similarly, on the Custom Low-Resolution Dataset, our technique demonstrated an accuracy of 0.8803, outpacing the closest competitor, Densenet, by approximately 0.5%.

In conclusion, this study aims to address the limitations of traditional sports video classification methods by proposing an improved deep learning model. Our research first introduces a method for selecting important video frames and applies a fuzzy noise attenuation algorithm to enhance video quality. Then, we use a dual-branch neural network to extract detailed features and classify videos through a densely connected neural network with specific activation functions. Our model was tested on a dataset of high-definition sports videos covering over 20 sports and a low-resolution dataset. This study emphasizes a robust approach to sports video classification, making key contributions to improving the processing capability for high-speed actions and low-resolution inputs, and ensuring superior performance at different resolutions. Our results pave the way for advancements in the field of sports video analysis and are significant for future applications in sports data analysis, content organization, and retrieval processes.

2. Methods

In this section, we begin by introducing the dataset used and describing the preprocessing steps. Next, we present our proposed model along with the training details and implementation specifics. Finally, we provide a summary of the evaluation metrics employed in our study.

2.1. DATA Set

The DeepSports-VDS is a meticulously curated dataset tailored for classification tasks in high-definition sports video content. It amalgamates videos sourced from HD sports broadcasts, high-speed cameras during live matches, and select segments from renowned datasets like UCF101 and Sports-1M (as shown in Figures 1 and 2). The primary selection criterion was superior video resolution and clarity, which positions this dataset as a benchmark for evaluating the influence of video quality on low-resolution sports video classification. Comprised of roughly 200 clips across 20 sports disciplines, each video has been rigorously annotated by expert sports professionals, ensuring precise label-content alignment. Designed specifically for category classification, clips are capped at 50 s, resulting in a dataset size of approximately 20 GB.

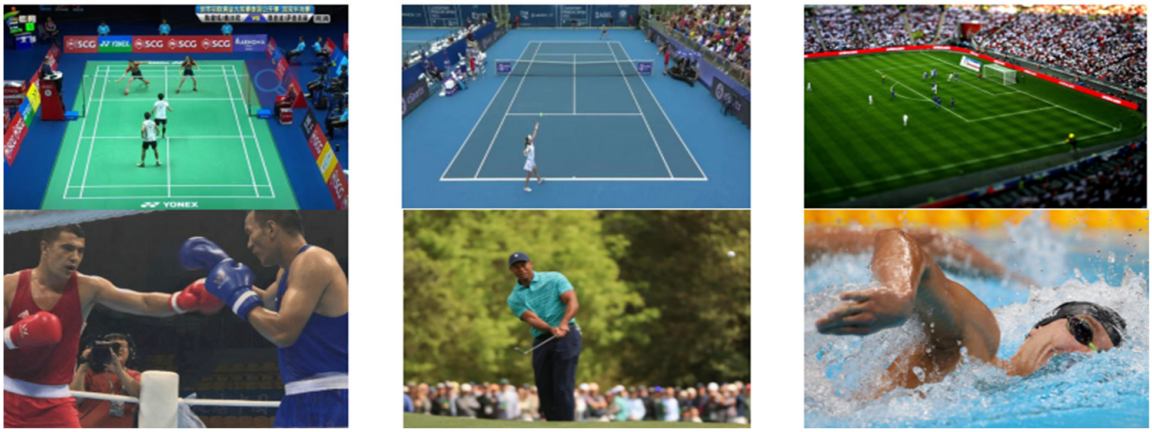

Figure 1. Image frames from a partial video dataset.

Figure 2. Images from self-collected sports video dataset (contains artifacts and low resolution).

In order to effectively train and evaluate our model, we have divided the DeepSports-VDS dataset into training, validation, and testing sets. Specifically, 80% of the dataset has been allocated to the training set, which will be used to train our model and fine-tune its parameters. The remaining 20% has been split evenly between the validation and testing sets, with each receiving 10% of the data. The validation set will be used to monitor the model's performance during training and make adjustments as needed, while the testing set will be used to evaluate the model's final performance on unseen data. This division ensures that our model is well-generalized and able to accurately classify sports videos from various sources and conditions.

To assess the performance of our proposed model across varied sports video qualities, we curated a low-resolution and blurred sports video dataset. This collection spans 20 prevalent sports disciplines, aggregated from online platforms and low-resolution cameras, mirroring real-world challenges inherent to genuine sports video processing. Our findings underscore the model's robustness in handling a spectrum of video qualities. The dataset comprises 100 video samples, post-preprocessing, each standardized to roughly 50 s. File sizes oscillate between 5 to 10 MB, with an average resolution nearing 480 P. Given the eclectic and genuine origins of the dataset, there is a pronounced variance in video quality, encompassing resolution, frame rate, clarity, noise, among other facets. Such disparities render our dataset more demanding and pragmatic, aptly reflecting the intricacies of processing authentic sports footage.

In collecting the dataset, we chose representative sports like basketball, soccer, and table tennis, ensuring its diversity and broad applicability. During labeling, we prioritized accuracy, employing professionals and emphasizing teamwork for consistent results.

2.2. Preprocessing

Given the inherent nature of sports videos, they are often marred by motion artifacts due to rapid movements. To address this, our study mandates preprocessing steps prior to classification, which encompasses artifact and noise elimination. Specifically, frames are extracted from the sports videos and resized to a uniform resolution of 520 × 520 pixels. We leverage the Laplacian second-order differential linear operator (as per Equation (1)) to discern high-frequency components in the images. A paucity of these components suggests potential blurriness and motion artifacts. By computing the image variance using the Laplacian operator filter, and setting a threshold at 0.3, we ascertain that images falling below this variance are likely blurred and artifact-laden.

$$\nabla f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x,y) \quad (1)$$

To tackle motion artifacts in images, one can infer that the pristine image has been displaced by L pixels, either horizontally or vertically. This displacement can be conceptualized as the original image being subjected to a Fourier transform with a blur kernel of magnitude L, oriented either horizontally or vertically. Given a camera exposure time of t, the object's displacement can be delineated into $x(t)$ and $y(t)$ for the horizontal and vertical axes, respectively. Thus, the motion-artifacted image can be derived by executing a Fourier transform on the original, as illustrated in Equation (2).

$$\begin{array}{l} G(x,y) = \int_0^T f(x - x(t), y - y(t))dt \\ g(x,y) = \int_0^T f(x - x(t), y - y(t))dt \end{array} \tag{2}$$

By applying the Fourier transform to $g(x,y)$ we obtain the formula:

$$\begin{array}{rl} G(u,v) & = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) e^{-j2\pi(ux+vy)} dxdy \\ & = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\int_0^t f(x - x(t), y - y(t))dt e^{-j2\pi(ux+vy)} dxdy \\ & = \int_0^t \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) e^{-j2\pi(ux+vy)} dxdy \end{array} \tag{3}$$

Assuming:

$$J(u,v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) e^{-j2\pi(ux+vy)} dxdy \tag{4}$$

$$k(u,v) = \int_0^t e^{-j2\pi(ux(t)+vy(t))} dt \tag{5}$$

Then, we can derive that

$$G(u,v) = J(u,v)k(u,v) \tag{6}$$

Here $G(u,v)$ represents the spectral amplitude of the transform. Utilizing the Radon transform, we ascertain the frequency offset angle, aligning with the blur kernel's orientation. The offset scale is derived from the spacing between consecutive wave patterns evident in the blurred image. Conclusively, executing the inverse Fourier transform facilitates the retrieval of a pristine image devoid of motion artifacts.

2.3. Model and Training

To extract the initial 80 frames from each video and mitigate potential noise from irrelevant backgrounds, we employ center cropping on these frames. The shorter dimension of the image dictates the side length of the central square, which is subsequently resized to 299×299 pixels via the OpenCV library. Consequently, each sports video input manifests as an $80 \times 299 \times 299 \times 3$ tensor.

In the first branch, the process begins with fine-grained feature extraction from the downsampled image, leading to the generation of Feature 1. This extraction, as detailed in Figure 3, starts by dividing the image into four blocks. Each block undergoes a specialized 1×1 convolution process, utilizing 32 filters. These filters are complemented by a ReLU activation function and a stride of 1, optimizing the extraction process. The output from each block is a feature map, which are then concatenated along the spatial dimension. This concatenated map undergoes a further transformation through a 1×1 convolution with 128 filters, same padding, and a stride of 1, effectively restoring the dimensions of the combined feature map. This restored map is then processed through two sequential Conv2D layers, each employing two 3×3 convolution kernels with a stride of 1 and 32 filters. These layers are followed by a 2D average pooling, a batch normalization layer to expedite convergence, and a Mish activation function layer, further refining the feature extraction. The process concludes with a downsampling layer, producing a refined feature map designated as f1.

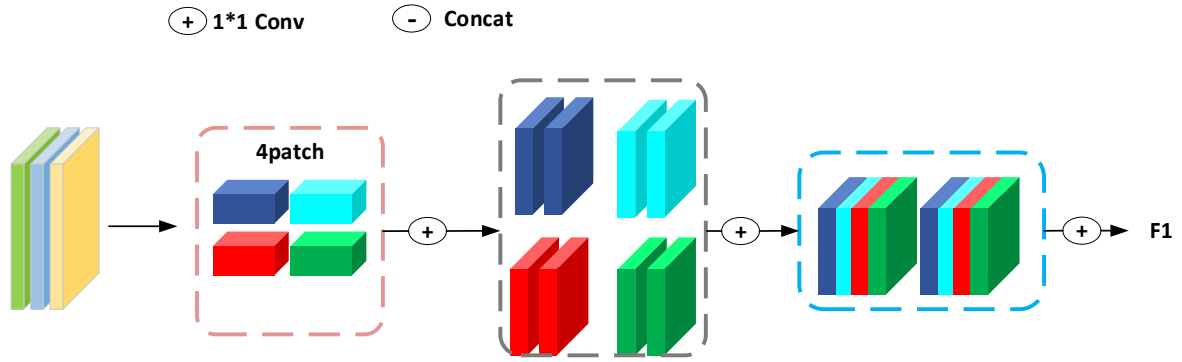

Figure 3. Fine−grained feature extraction architecture.

The second branch initiates with F0 undergoing feature extraction through two Conv2D layers, followed by a downsampling step to obtain feature map F2, as detailed in Figure 4. Subsequently, the network pathway behind F2 splits into a main branch and an auxiliary branch. In the −main branch, F2 is processed through a fine-grained feature layer, transforming it into F2′. This new feature map, F2′, is then concatenated with F1 in the channel dimension and further processed through a Conv2D layer, resulting in F2″. Concurrently, in the auxiliary branch, F2 first passes through a Conv2D layer, followed by a downsampling layer. After skip concatenation with F1, it undergoes another round of fine-grained feature extraction. This is followed by an upsampling layer employing bilinear interpolation, which yields feature F3. Finally, F3 is concatenated with F2″, culminating in the final feature, F_final. This intricate process, where features from different convolutional depths are fused, encapsulates a richer array of local and global information. This comprehensive fusion significantly enhances the model's capability to understand and interpret images, leveraging the unique strengths of each convolutional layer and processing step.

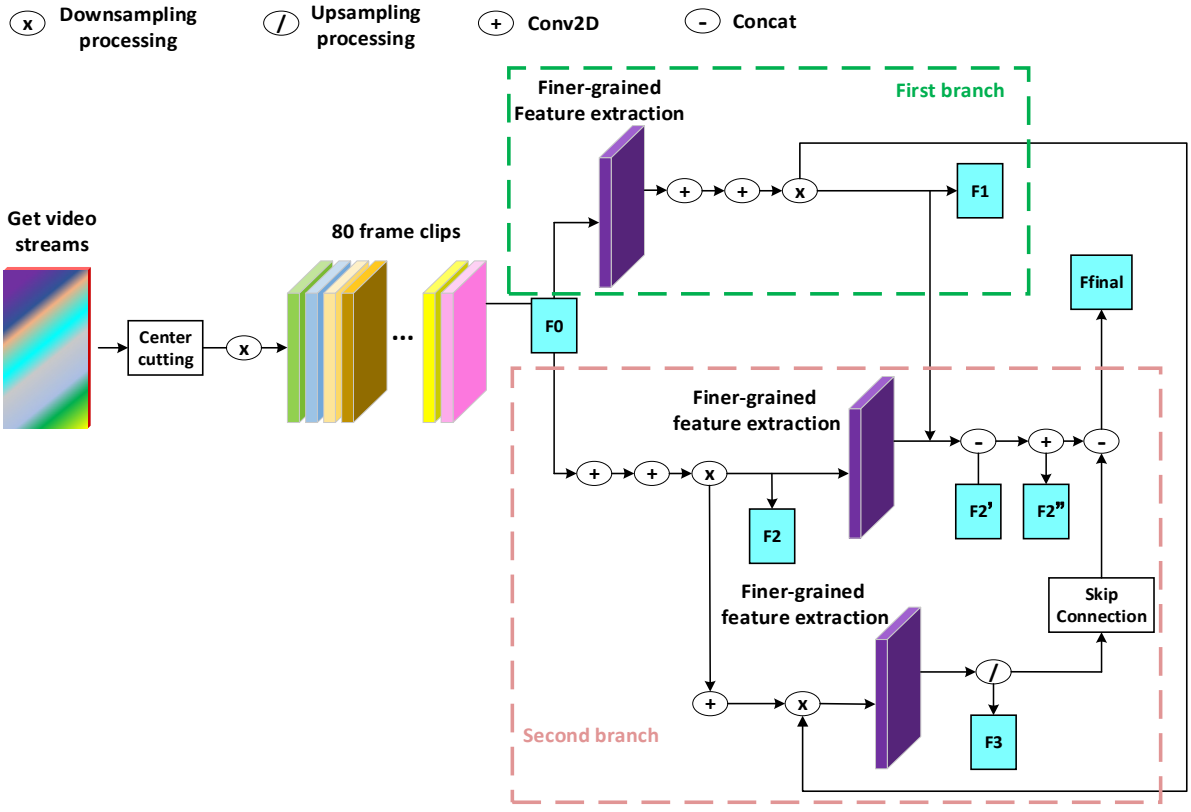

Figure 4. Two−branch Neural network model architecture for sports video classification.

The specific classification mechanism is as follows:

The latter segment of our methodology zeroes in on feature classification. Leveraging the backbone network, we extract pivotal features from 80 curated frames per video. Given the multi-dimensional character of these features, a Flatten layer is employed to reshape them into a linear format, ensuring seamless interfacing with the ensuing network layers and facilitating efficient data flow.

For the classification module, we harness the prowess of an advanced LSTM network adept at serializing image frames. Videos, by nature, are sequential, demanding a distinct analytical approach compared to standalone images. LSTMs, with their capacity to store and process extended sequences, are paramount for discerning the temporal intricacies inherent to video data.

To bolster the LSTM's sequential comprehension, position encoding is integrated into each frame. This encoding, acting as a sequential identifier, guarantees the LSTM's adherence to the original frame order, forestalling potential data misalignment that could compromise classification accuracy.

Post-encoding, the data is channeled into the LSTM, architected to manage sequential constructs. Within this network, layers are orchestrated such that one layer's hidden state cascades as input to its successor. This layered interplay captures nuanced temporal dynamics, ensuring the LSTM's output encapsulates a holistic video sequence representation, priming it for precise classification.

To gauge our model's efficacy, we deploy a binary cross-entropy loss function during training. This metric quantifies the divergence between predicted outcomes and actual labels. By fine-tuning model parameters via the gradient descent technique, our aim is to create a model epitomizing minimal error.

Subsequently, the feature extraction backbone and LSTM converge, culminating in an integrated model adept at end-to-end tasks. This holistic model ingests raw video frames, navigates them through the intricate network architecture, and ultimately yields classification verdicts. To fortify model resilience and curb overfitting, we incorporate strategies like dropout—for diversified feature dependency—and early stopping, which halts training upon discerning peak performance on a validation set.

2.4. Implementation Details

In this work, our framework was implemented using the PyTorch library and utilized two NVIDIA GeForce RTX 3080 GPUs (NVIDIA, California, USA) for computation. We set the batch size to 16 and used the Adam optimizer for training. The number of training epochs was configured to 200. To prevent overfitting, early stopping was employed. Additionally, we incorporated a dynamic learning rate adjustment using the cosine annealing algorithm. The learning rate was set to vary within the range of 0.001 to 0.00001, allowing for efficient navigation through the model's loss landscape and aiding in the convergence to an optimal set of weights.

2.5. Model Evaluation

In the sports video classification task, it was essential to evaluate the performance of the model to ensure its effectiveness and reliability. Typically, accuracy, precision, recall, F1-score, and confusion matrix are used to assess the performance of the classification model.

3. Results

In this section, we present examples of video data classification using the model proposed in our study (as shown in Figure 5). We demonstrate the classification results for inputs using both high-resolution and low-resolution datasets, and compare these with other traditional models. Finally, we use a Confusion Matrix to illustrate the strong generalization capabilities of our classification algorithms.

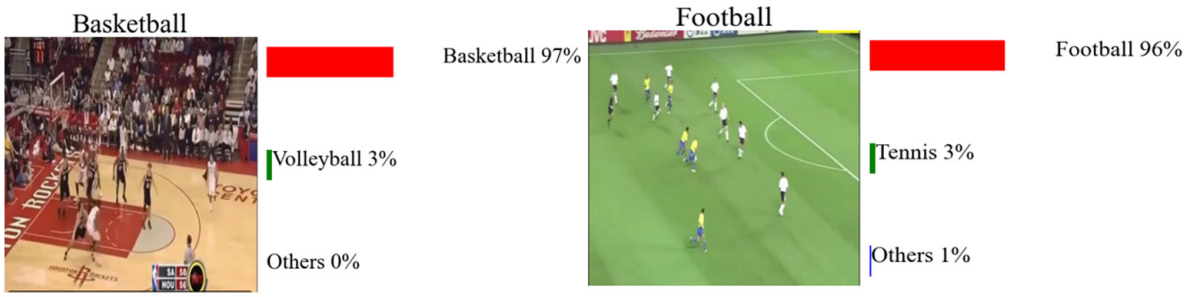

Figure 5. Example of classification results of proposed algorithm.

In this study, we compared our proposed algorithm with other renowned models in the SPORTS1 m and a bespoke dataset, which encompass sports videos of varying resolutions. Based on this, we elucidated the effectiveness of our proposed algorithm. Initially, the evaluation was conducted on the SPORTS1 m dataset. The corresponding results can be referred to in Table 1 and Figure 6.

Table 1. Classification results of different classification algorithms on the high-resolution dataset.

Model	Precision	Accuracy	F-Score	Recall
DenseNet	0.9690	0.9730	0.9670	0.9651
VggNet-16	0.9338	0.9631	0.9238	0.9140
Inception v3	0.9531	0.9781	0.9654	0.9677
ResNet-50	0.9380	0.9645	0.9528	0.9681
Proposed method	0.9718	0.9804	0.9761	0.9723

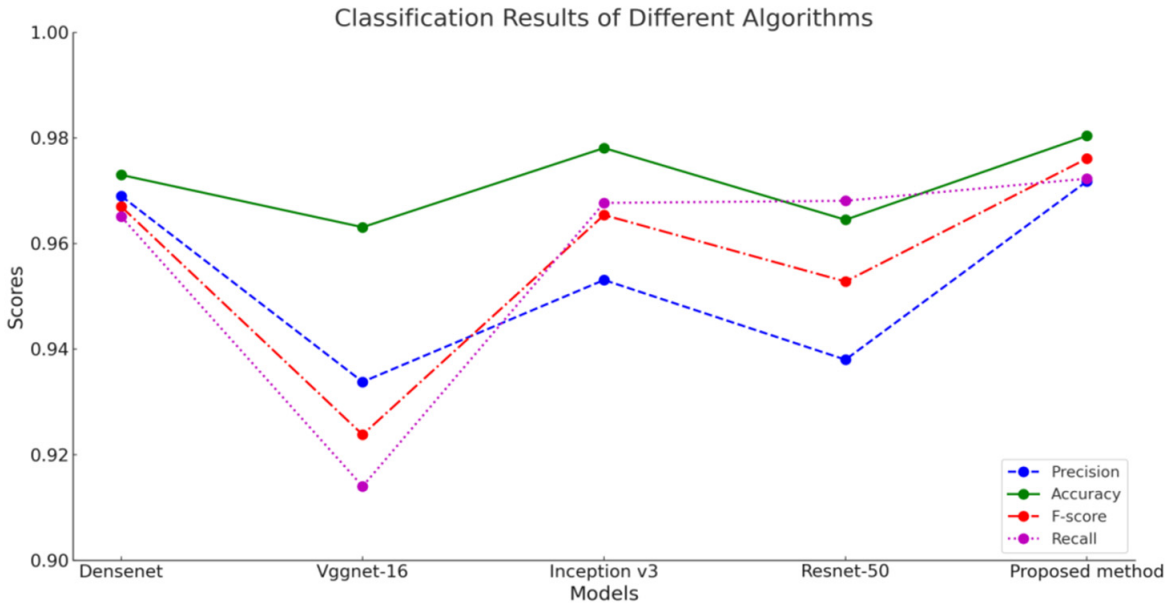

Figure 6. Accuracy and loss curves of classification on the high-resolution dataset.

In light of the empirical findings presented, it is evident that our proposed algorithm demonstrates superior performance in comparison to alternative models when evaluated on a high-resolution dataset. Specifically, the method achieved a precision of 0.9718, an accuracy rate of 0.9804, an F-score of 0.9761, and a recall rate of 0.9723.

Next, we evaluate the models on a Low-Resolution Dataset, specifically curated for sports video classification. The results are as follows Table 2, Figures 6 and 7.

Table 2. Classification results of different classification algorithms on the low-resolution dataset.

Model	Precision	Accuracy	F-Score	Recall
DenseNet	0.8627	0.8763	0.8675	0.8592
VggNet-16	0.8532	0.8604	0.8561	0.8523
Inception v3	0.8571	0.8682	0.8634	0.8605
ResNet-50	0.8583	0.8651	0.8602	0.8574
Proposed method	0.8725	0.8803	0.8752	0.8701

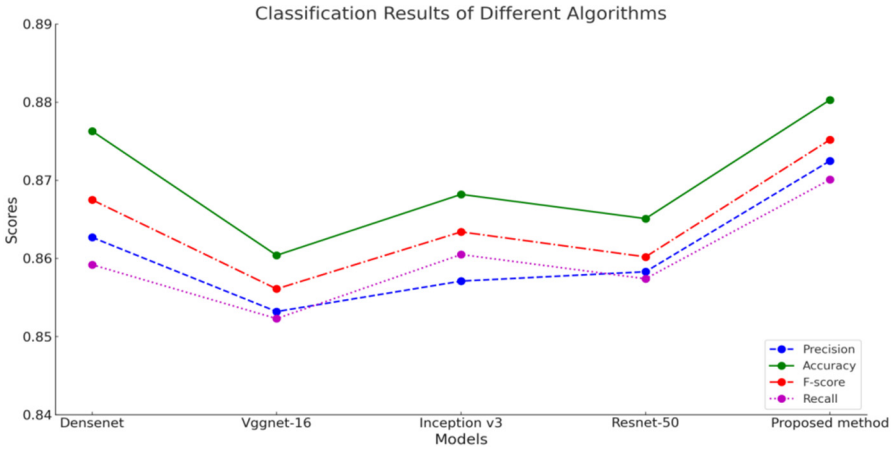

Figure 7. Accuracy and loss curves of classification on the low-resolution dataset.

Based on the empirical data presented, it is manifestly evident that our proposed algorithm surpasses its peers when evaluated on the low-resolution dataset, excelling in metrics such as precision, accuracy, F-score, and recall. Concurrently, the Confusion Matrix derived from the Validation Set accentuates the robust generalization prowess of our model (as shown in Figure 8), potentially attributable to the incorporation of the deblurring algorithm.

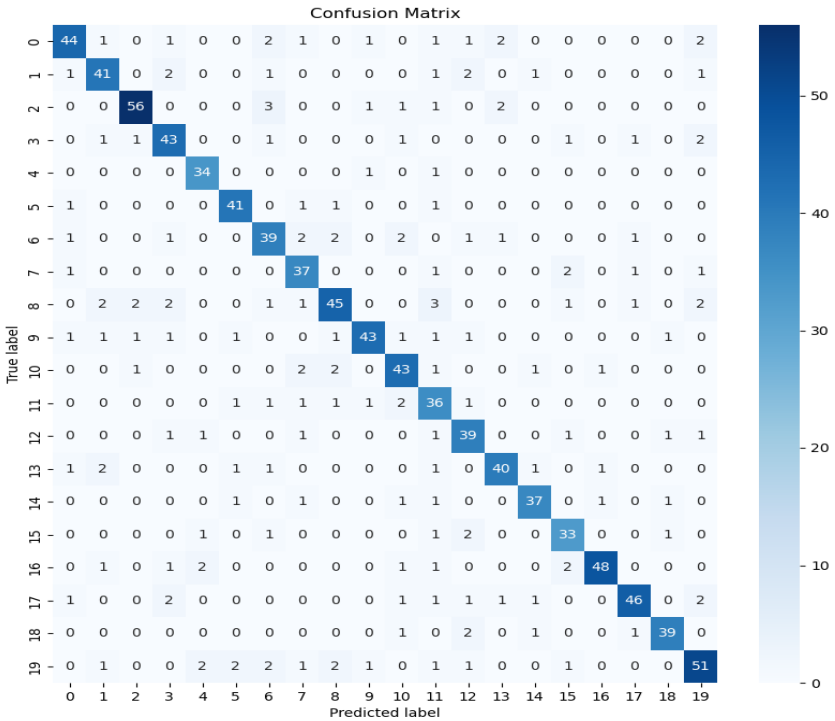

Figure 8. Confusion Matrix under the Validation Set of the low-resolution dataset.

In conclusion, the empirical evidence underscores that our proposed sports video classification algorithm consistently outshines its contemporaneous, well-regarded counterparts across both the SPORTS1 m and bespoke datasets. Such findings unequivocally attest to the robustness and efficacy of our model in the realm of sports video classification.

4. Discussion

In the realm of artificial intelligence and deep learning, the classification of sports videos remains a pivotal challenge. In our investigation, we endeavor to elucidate the efficacy of deep learning methodologies in the categorization of sports footage. We introduce an avant-garde model tailored for sports video taxonomy and juxtapose its prowess against renowned architectures, namely DenseNet, VGGNet-16, Inception V3, and ResNet-50. Our paradigm encompasses the preprocessing of video frames, feature extraction via a bifurcated neural network architecture, and subsequent classification leveraging an augmented LSTM framework.

Our empirical results accentuate the preeminence of our proposed algorithm vis-à-vis its counterparts. The model manifested a precision of 0.9718, accuracy of 0.9804, F-score of 0.9761, and recall of 0.9723, outstripping the aforementioned architectures across multifarious performance indices. Notably, within our bespoke dataset, the model attained an accuracy of 87.25% for low-resolution sports videos and those imbued with artifacts, underscoring its potential ramifications in diverse sports-centric applications.

Given its superior efficacy in sports video taxonomy, our model stands poised to confer substantial advantages to myriad stakeholders within the sports milieu. Potential applications encompass automated annotation and indexing of sports footage, thereby facilitating seamless content retrieval for coaches, athletes, and aficionados. Furthermore, the algorithm can be harnessed for the synthesis of video digests, offering spectators a succinct rendition of athletic events.

Our innovation also holds promise in the following arenas:

(i) Injury Prophylaxis and Rehabilitation: Video studies can be applied for sports injury analysis [17]. Thus, the model can discern potentially perilous or flawed movements predisposing to injuries. Additionally, real-time detection empowers coaches to intercede with corrective guidance [18], attenuating injury risk. Adaptations of the model can also serve rehabilitative ends, e.g., guiding injured athletes in their recuperative journey, as the recognized injury patterns have the potential to guide injury prevention [19]. (ii) Augmented Fan Engagement: Our model can curate captivating content for sports enthusiasts [20], autonomously crafting highlights [21], statistics [22], and analytical insights [23]. This augments the spectator experience, furnishing a profound comprehension of the sport and its participants. Integration into sports streaming platforms can also offer tailored content suggestions, resonating with users' predilections. (iii) Sports Pedagogy and Training: The model can be integrated into educational frameworks, enlightening students and budding athletes about techniques, tactics, and best practices across sports disciplines. Automated video analysis [24,25] can engender a more immersive pedagogical experience.

Our research initially pre-trains the neural network on a large sports video dataset and then fine-tunes the network for specific sports or tasks. This approach aims to explore the potential of transfer learning, which can improve model performance and reduce the need for extensive domain expertise. In the era of large models, our findings contribute to enhancing the quality of data for precise training of 'modules' based on a transferable parameter system established from vast general-category data. This, in turn, provides valuable support for optimizing the visual modality branch of AGI (Artificial General Intelligence).

To encapsulate, the sports video taxonomy model delineated herein holds immense promise in revolutionizing the sports sector, offering invaluable insights and automating a plethora of tasks. Its versatility spans performance analytics, injury mitigation, talent identification, and fan engagement. Future endeavors can further refine its capabilities and extend its applicability, rendering it an indispensable asset in the sports domain.

Our research also has some limitations. High computational requirements make deployment difficult on embedded systems and edge devices. Efficient model designs are being explored [26]. By integrating the research in this field, the robustness, security, and reliability of real-world systems adopting sports video classification methods can be enhanced. The precision and accuracy of this model in classifying videos still have a gap compared to 100%. The most direct reason for the classification errors is the small amount of data, which is also highly heterogeneous. Additionally, in low-resolution videos, certain sports like badminton, tennis, and volleyball have minimal feature differences, which can contribute to classification errors. In future work, this can be addressed by increasing the sample size of the data, performing data augmentation on such data, or using pre-training and transfer learning on this type of data. Moreover, considering the use of transformer architectures to enhance the contextual connections of videos could also improve classification accuracy.

5. Conclusions

This study proposes a novel deep learning framework specifically designed for sports video classification. Addressing the limitations of conventional methods in dealing with fast-paced and low-resolution videos, we adopt techniques such as key frame selection, fuzzy noise reduction algorithm, and dual-branch neural network. Experimental results demonstrate the superior performance of our method on two different sports video datasets, outperforming other renowned baselines. This suggests that our approach possesses high accuracy and robustness in handling sports video classification tasks under various conditions. The results of this study will also be practically applied in sports training and video analysis.

Author Contributions: Conceptualization, T.G., M.Z. and W.L.; methodology, T.G. and M.Z.; software, M.Z., Y.Z. (Yifan Zhu) and J.Y.; validation, W.L. and J.Y.; formal analysis, T.G. and M.Z.; investigation, X.P.; resources, Y.Z. (Youjian Zhang); data curation, M.Z.; writing—original draft preparation, T.G. and M.Z.; writing—review and editing, Y.Z. (Yifan Zhu), Y.Z. (Youjian Zhang) and W.L.; visualization, T.G. and Y.Z. (Yifan Zhu); supervision, W.L.; project administration, J.Y. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The high-definition DeepSports-VDS dataset, featuring 200 clips across 20 sports disciplines, and the low-resolution sports video dataset comprising 100 samples are accessible for research purposes upon direct request from the authors. These datasets authentically replicate genuine sports video challenges, encompassing diverse qualities mirroring real-world conditions. Requests for access should be directed to the corresponding authors in line with established academic protocols.

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

References

1. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3D residual networks. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 5534–5542.
2. Bagautdinov, T.M.; Alahi, A.; Fleuret, F.; Fua, P.; Savarese, S. Social scene understanding: End-to-End multi-person action localization and collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 3425–3434.
3. Tang, Y.; Wang, Z.; Li, P.; Lu, J.; Yang, M.; Zhou, J. Mining semantics-preserving attention for group activity recognition. In Proceedings of the 26th ACM International Conference on Multimedia Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1283–1291.
4. Cao, S.; Wang, B.; Zhang, W.; Ma, L. Visual consensus modeling for video-text retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; pp. 167–175.
5. Yu, H.; Cheng, S.; Ni, B.; Wang, M.; Zhang, J.; Yang, X. Fine-Grained video captioning for sports narrative. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6006–6015.

6. Arnold, R.; Fletcher, D.; Molyneux, L. Performance leadership and management in elite sport: Recommendations, advice and suggestions from national performance directors. *Eur. Sport Manag. Q.* **2012**, *12*, 317–336. [CrossRef]
7. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, 23–28 June 2008.
8. Zebhi, S.; Al-Modarresi, S.M.T.; Abootalebi, V. Converting video classification problem to imageclassification with global descriptors and pre-trained network. *IET Comput. Vis.* **2020**, *14*, 614–624. [CrossRef]
9. Cust, E.; Sweeting, A.J.; Ball, K.; Robertson, S. Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *J. Sports Sci.* **2019**, *37*, 568–600. [CrossRef] [PubMed]
10. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
11. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, 23–28 June 2008.
12. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-Style ConvNets Great Again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13733–13742.
13. Lin, H.; Jegelka, S. ResNet with one-neuron hidden layers is a Universal Approximator. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6172–6181.
14. Mahdi, M.; Yunpeng, Z.; Guoning, C. Resource allocation in cloud computing using genetic algorithm and neural network. In Proceedings of the 2023 IEEE 8th International Conference on Smart Cloud (SmartCloud), Tokyo, Japan, 16–18 September 2023; pp. 25–32.
15. Rasul, C.; Ata, Z. A deep neural network modeling methodology for efficient EMC assessment of shielding enclosures using MECA-generated RCS training data. *IEEE Trans. Electromagn. Compat.* **2023**, *65*, 1782–1792.
16. Xiaoping, G. Intelligent Sports Video Classification Based on Deep Neural Network (DNN) Algorithm and Transfer Learning. *Comput. Intell. Neurosci.* **2021**, *2021*, 1825273.
17. Sugimoto, D.; Myer, G.; Micheli, L.; Hewett, T. ABCs of Evidence-based anterior cruciate ligament injury prevention strategies in female athletes. *Curr. Phys. Med. Rehabil. Rep.* **2015**, *3*, 43–49. [CrossRef]
18. Chatzitofis, A.; Zarpalas, D.; Daras, P. A computerized system for real-time exercise performance monitoring and e-coaching using motion capture data. In Proceedings of the Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18–21 November 2017; pp. 243–247.
19. Klein, C.; Luig, P.; Henke, T.; Bloch, H.; Platen, P. Nine typical injury patterns in German professional male football (soccer): A systematic visual video analysis of 345 match injuries. *Br. J. Sports Med.* **2021**, *55*, 390–396. [CrossRef]
20. Hazari, S. Investigating social media consumption, sports enthusiasm, and gender on sponsorship outcomes in the context of Rio Olympics. *Int. J. Sports Mark. Spons.* **2018**, *19*, 396–414. [CrossRef]
21. Stride, A.; Fitzgerald, H.; Allison, W. A narrative approach: The possibilities for sport management. *Sport Manag. Rev.* **2017**, *20*, 33–42. [CrossRef]
22. Soomro, K.; Zamir, A. Action recognition in realistic sports videos. In *Advances in Computer Vision and Pattern Recognition*; Springer: Cham, Switzerland, 2014; pp. 181–208.
23. Stein, M.; Janetzko, H.; Lamprecht, A.; Breitkreutz, T.; Zimmermann, P.; Goldlücke, B.; Keim, D. Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE Trans. Vis. Comput. Graph.* **2017**, *24*, 13–22. [CrossRef] [PubMed]
24. Thomas, G.; Gade, R.; Moeslund, T.; Carr, P.; Hilton, A. Computer vision for sports: Current applications and research topics. *Comput. Vis. Image Underst.* **2017**, *159*, 3–18. [CrossRef]
25. Voeikov, R.; Falaleev, N.; Baikulov, R. TTNet: Real-time temporal and spatial video analysis of table tennis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Nashville, TN, USA, 19–25 June 2021; pp. 884–885.
26. AlSobeh, A.M.; Magableh, A.A. BlockASP: A Framework for AOP-Based Model Checking Blockchain System. *IEEE Access* **2023**, *11*, 115062–115075. [CrossRef]