IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 1039-1053
Journal homepage: http://iieta.org/journals/ts

Instance Segmentation and Fine-Grained Contour Extraction of Basketball Players from Videos under Complex Occlusion Scenarios

Delong Jia

Department of Physical Education, Shandong Technology and Business University, Yantai 264005, China

Corresponding Author Email: 19153509328@163.com

https://doi.org/10.18280/ts.430238

Received: 12 November 2025
Revised: 27 February 2026
Accepted: 9 March 2026
Available online: 30 April 2026

Keywords:
basketball video image segmentation, complex occlusion, Dynamic Snake Convolution, Elliptic Fourier Descriptors, Signed Distance Function, contour refinement, multi-scale feature fusion

ABSTRACT

In basketball game videos, challenges such as frequent limb bending, dense multi-player occlusion, and complex court lighting conditions severely degrade the performance of conventional instance segmentation methods, often resulting in coarse mask boundaries, distorted contours in occluded regions, and insufficient representation accuracy. To address these issues, this paper proposes a unified framework that integrates dynamic deformable feature extraction, parametric shape encoding, and continuous distance field optimization for precise basketball player instance segmentation and contour refinement. First, a boundary-enhanced multi-scale feature extraction network based on Dynamic Snake Convolution (DSC) is constructed to accurately capture slender and curved limb structures while enabling bidirectional fusion of semantic and edge features. Second, departing from the traditional binary discrete mask prediction paradigm, a two-stage continuous contour prediction mechanism is introduced, which combines Elliptic Fourier Descriptors for global shape encoding with a Signed Distance Function (SDF) field for local boundary refinement, achieving compact shape representation and sub-pixel-level contour restoration. Furthermore, a depth-hierarchy-aware copy-paste data augmentation strategy with spatial constraints is designed to simulate realistic occlusion scenarios, and a level-set evolution-based active contour algorithm is incorporated for post-processing refinement in heavily occluded areas. Finally, a multi-task joint weighted loss function is formulated to enable collaborative optimization across multiple branches. Extensive experiments conducted on both a self-constructed basketball occlusion dataset and publicly available sports vision benchmarks demonstrate that the proposed method significantly outperforms state-of-the-art instance segmentation approaches in terms of segmentation accuracy and contour fitting quality under severe occlusion conditions. The proposed approach effectively accomplishes accurate and smooth player instance segmentation, providing reliable visual technical support for practical applications such as intelligent tactical analysis, quantitative sports pose assessment, and athlete behavior recognition.

1. INTRODUCTION

With the rapid iteration of computer vision technologies [1] and their deep integration with the smart sports industry, intelligent sports event analysis has become a core driving force for the digital transformation of the sports industry [2, 3]. As one of the most globally popular competitive sports, basketball game analysis is gradually upgrading from traditional manual statistics to an intelligent and refined direction, and the core supporting role of precise visual perception of athletes is becoming increasingly prominent. However, the dynamic and complex environment of basketball courts brings many intractable technical challenges to athlete instance segmentation and contour extraction [4, 5]: athletes' limbs exhibit slender and curved non-rigid shapes, and actions such as torso twisting and limb stretching during movement lead to constant changes in target morphology; in high-intensity confrontation scenarios, densely arranged players are prone to large-scale limb cross-occlusion, with some areas even experiencing complete occlusion, resulting in incomplete target features; fluctuations in court lighting intensity, motion artifacts, and blurring effects caused by fast motion further exacerbate image quality fluctuations; meanwhile, there are significant scale differences between athletes, referees, and coaches on the court, coupled with rapid displacement, further increasing the difficulty of feature extraction and contour localization. As the underlying core technology of intelligent basketball game analysis, the performance of instance segmentation and precise contour extraction directly determines the accuracy and reliability of upper-level applications such as player trajectory tracking, quantitative action scoring, tactical decomposition, and intelligent event commentary. Therefore, conducting research on athlete instance segmentation and contour refinement for complex occlusion scenarios in basketball game videos holds significant practical necessity [6-8]. From a theoretical perspective, this study can improve the theoretical system of continuous contour representation for non-rigid human targets

under complex dynamic occlusion scenarios, break through the accuracy bottleneck of traditional discrete mask segmentation, and establish a new visual segmentation paradigm integrating deformable convolution feature extraction and parametric shape priors [9, 10], enriching the research content and technical paths in the field of sports-specific visual image processing. From an application perspective, the research results can be directly implemented in professional basketball game intelligent analysis systems [11], providing coaching teams with accurate tactical optimization bases; they can be applied to campus basketball teaching correction [12], assisting students in standardizing movement postures; they can also be extended to scenarios such as public fitness posture assessment and sports big data intelligent analysis [13, 14], possessing extremely high engineering implementation value and industrial promotion potential.

Currently, scholars at home and abroad have conducted extensive research on instance segmentation, sports human segmentation, contour extraction, and occlusion optimization, forming a series of research results. However, specific research targeting complex occlusion scenarios in basketball game videos still suffers from many common defects, making it difficult to meet actual application demands. Mainstream instance segmentation models rely on pixel-level binary mask output; their inherent discrete nature leads to obvious stepped aliasing on segmentation edges, making sub-pixel level contour extraction impossible [15, 16]. In areas where athletes' limbs cross and occlude, pixel misclassification is prone to occur, leading to contour boundary distortion. Traditional deformable convolutions adopt an independent sampling point offset mechanism lacking continuous path constraints, making them unsuitable for adapting to the complex geometric shapes of basketball players' limb bending and torso twisting. Their integrity in extracting long-distance limb features is insufficient, leading to feature fracture problems. Existing contour extraction methods mostly focus on local edge detection [17, 18], lacking global human shape prior constraints. In areas where occlusion leads to missing contours, contour deformation and limb fracture are highly likely to occur, making it impossible to achieve accurate restoration of complete contours. General image data augmentation strategies do not consider the spatial depth logic of the court and cannot simulate the real physical occlusion relationship of "near occluding far." Synthetic training samples differ greatly from actual court scenes, resulting in insufficient generalization capability of models in complex occlusion scenarios. Furthermore, existing segmentation frameworks lack dedicated loss constraints for sports human targets [19]; the optimization weight distribution for edge features and shape features is unreasonable, and the collaborative fitting effect of multi-task branches is poor, making it difficult to balance segmentation accuracy and contour refinement [20]. The existence of these problems severely restricts the application and promotion of athlete instance segmentation and contour extraction technologies in intelligent basketball game analysis, necessitating the proposal of a targeted solution.

Aiming at the above research deficiencies, this paper conducts research on athlete instance segmentation and contour refinement for complex occlusion scenarios in basketball game videos. The core innovations and contributions are as follows: designing a boundary-enhanced multi-scale Dynamic Snake Convolution (DSC) feature extraction backbone network, introducing cumulative offset constraints to match athlete limb features, and realizing deep fusion of semantic and edge features; proposing a dual-stage continuous contour prediction framework combining Elliptic Fourier global shape encoding and Signed Distance Function (SDF) local refinement to achieve compact contour representation and completion in occluded areas; constructing a position-aware depth-constrained copy-paste data augmentation method to improve model occlusion robustness; designing a multi-constraint fusion active contour post-processing mechanism to optimize contour refinement; building a multi-task joint weighted loss function to achieve collaborative optimization of all branch tasks.

The organizational structure of the subsequent chapters of this paper is as follows: Chapter 2 elaborates in detail on the athlete instance segmentation and contour refinement method proposed in this paper, including the specific design of the feature extraction network, the dual-stage contour prediction mechanism, the occlusion-aware training strategy, and the multi-task joint optimization loss function; Chapter 3 verifies the effectiveness and superiority of the proposed method through multiple sets of comparative experiments and ablation experiments, and provides a detailed analysis and discussion of the experimental results; Chapter 4 summarizes the full text's research work, objectively analyzes the limitations of the research, and looks forward to future research directions.

2. ATHLETE SEGMENTATION AND CONTOUR REFINEMENT METHOD FOR BASKETBALL OCCLUSION SCENARIOS

2.1 Boundary-enhanced multi-scale Dynamic Snake Convolution feature extraction backbone network

This chapter is guided by the characteristics of non-rigid deformation and dense occlusion of athletes on basketball courts, constructing a feature extraction backbone architecture adapted to the human body limb morphology. The network uses Residual Network with 50 Layers (ResNet-50) and Residual Network with 101 Layers (ResNet-101) as the basic feature encoding base, reconstructs and replaces all standard convolutional structures within the network conv2x to conv5x levels, and introduces DSC to complete the global upgrade of the feature sampling method. The overall architecture follows the hierarchical design idea of basic feature extraction, multi-scale feature fusion, and edge perception enhancement, relying on the residual structure to complete the layer-by-layer mapping from shallow image textures to deep semantics, while combining the Feature Pyramid Network to construct a multi-resolution feature representation system. The network sets four feature output levels, corresponding to feature maps with different downsampling rates. Each level independently completes feature encoding and information interaction, and adds bypass branches to construct a parallel representation structure for semantics and edges. The model uniformly receives game video frames with a resolution of 800×1280 as input, and outputs multi-scale feature maps with both semantic expression and edge details after backbone network encoding, providing a stable feature input basis for subsequent instance localization and contour fitting tasks. Figure 1 shows the overall framework diagram of athlete segmentation and contour extraction for basketball occlusion scenarios.

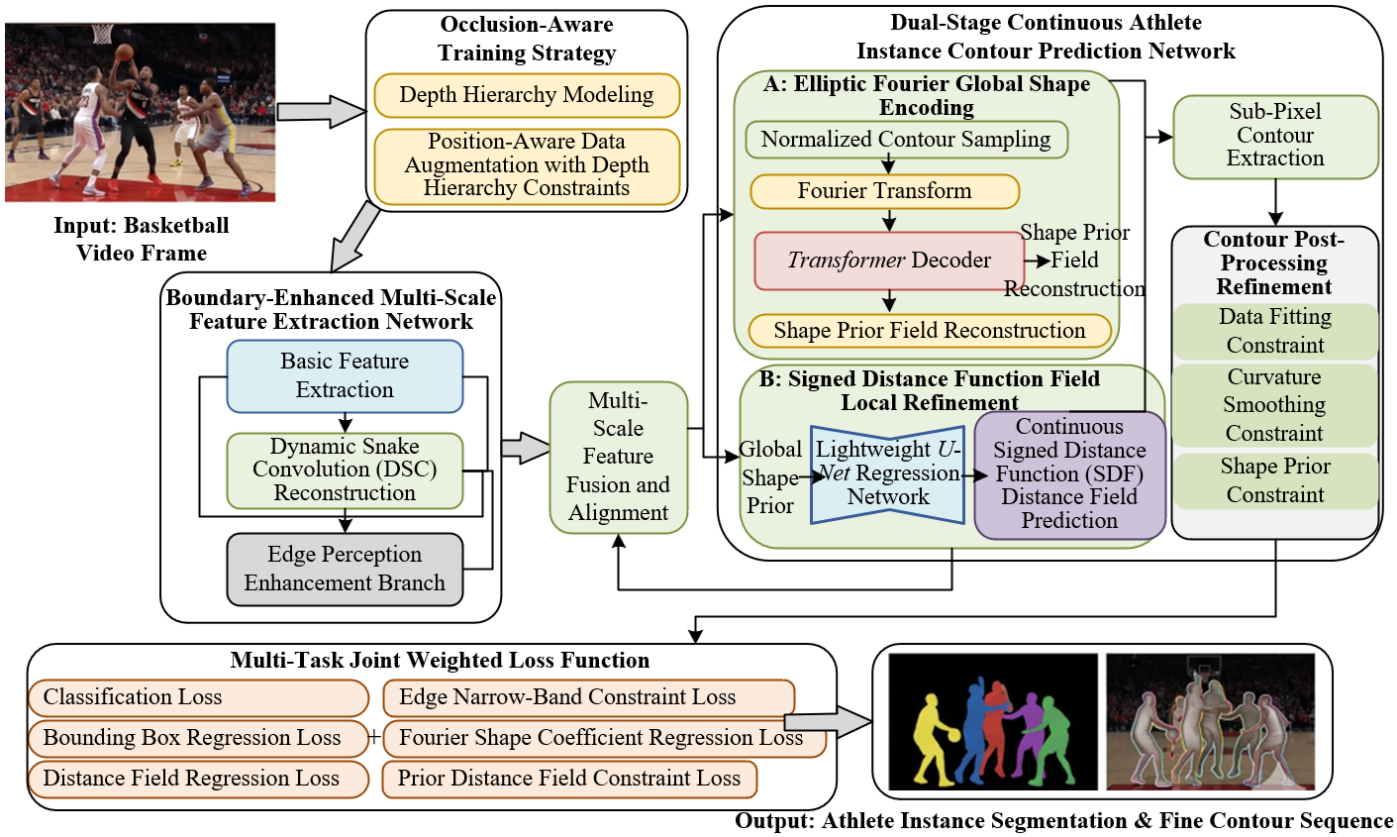

Figure 1. Overall framework diagram of athlete segmentation and contour extraction for basketball occlusion scenarios

As the core encoding unit of the backbone network, DSC redefines the convolution kernel spatial sampling pattern to adapt to the geometric shapes of human limb bending and torso twisting. The module follows the basic receptive field configuration of 3 × 3 convolution, discretizes the convolution neighborhood into nine independent sampling nodes, each node can learn a two-dimensional spatial offset to adaptively fit the target contour trend. The spatial positions of the sampling nodes no longer adopt an independent update mode but establish a cumulative transfer relationship of offsets between neighboring nodes. The calculation method of node positions can be expressed as:

$$p_i=p_0+\Delta_i+\sum_{j=1}^{i-1}\Delta_j \tag{1}$$

where, p_0 is the convolution center coordinate, and Δ_i represents the two-dimensional offset vector of the i-th sampling point. The offset is predicted adaptively by a separable convolution structure, capturing local spatial correlations through depth-wise convolution, and then compressing the channel dimension via pointwise convolution to output offset parameters. The cumulative transfer characteristic of node offsets enables the entire sampling sequence to form a continuous smooth curve trajectory, effectively solving the problem of discrete fracture in the sampling path of traditional deformable convolution, and achieving stable capture of complete features of slender limb structures.

To control the reasonable distribution range of sampling offsets and enhance feature responses in key regions, this paper introduces regularization constraints and channel adaptive enhancement mechanisms to DSC. Figure 2 shows the schematic diagram of the boundary-enhanced multi-scale DSC feature extraction module. Aiming at the problem of sampling region out-of-bounds caused by excessive offset vectors, L2 norm constraints are used to regulate the offset amplitude. The constraint loss expression is:

$$R(\Delta)=\lambda_\Delta\sum_{i=1}^{9}\|\Delta_i\|_2^2 \tag{2}$$

where, $R(\Delta)$ is the offset regularization loss function, λ_Δ is the regularization balance coefficient, set to 0.01 in this paper; $\|\Delta_i\|_2^2$ represents the L2 norm of the i-th sampling point offset vector, and the summation range covers all nine sampling nodes in the convolution neighborhood. This constraint can synchronously optimize the convolution weights and offset parameters during backpropagation, limiting the irregular growth of offsets. Meanwhile, a single-pixel offset threshold is set to prevent sampling nodes from exceeding the effective range of the convolution neighborhood, ensuring the physical meaning of convolution operations and the validity of feature sampling.

The regularization coefficient is set to 0.01, which synchronously optimizes the convolution weights and offset parameters during model backpropagation, while limiting the offset pixel threshold of a single node to maintain the effective operation range of the convolution neighborhood. After the convolution feature encoding is completed, a channel attention structure is connected, which aggregates the global feature information of each channel through global average pooling, relies on a two-layer fully connected layer to complete channel importance weight modeling, and uses an activation function to complete weight normalization mapping. After the feature map and channel weights are fused element-wise, it can automatically enhance the feature response of key regions such as limb joints and contour edges, and weaken the feature interference caused by court backgrounds and light shadows.

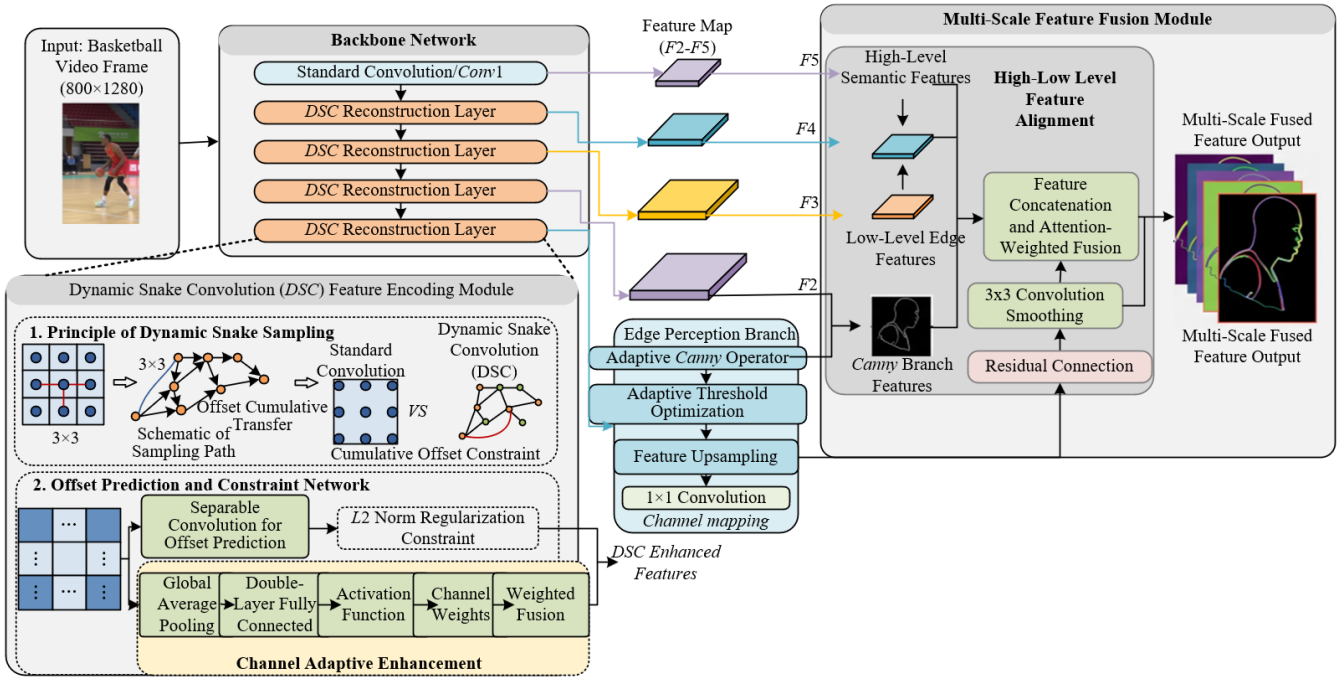

Figure 2. Schematic diagram of the boundary-enhanced multi-scale Dynamic Snake Convolution (DSC) feature extraction module

The inherent resolution differences and information representation biases of multi-scale feature levels easily cause representation defects such as blurred boundaries in high-level semantic features and missing semantics in low-level edge features. Therefore, this paper constructs independent edge perception branches for each feature level, relying on the adaptive Canny operator to complete low-level contour information extraction, and the algorithm threshold can be dynamically iterated and optimized with model training. The edge branch output features undergo 1 × 1 convolution to complete channel dimension mapping, keeping them consistent with the same-level backbone semantic features, achieving dimension matching of multi-branch features. Aiming at the problem of inconsistent resolution of high- and low-level features, bilinear interpolation is used to complete bidirectional alignment of feature scales: low-level edge features are upsampled to match high-level semantic scales, and high-level semantic features are downsampled to adapt to low-level detail scales, establishing a bidirectional flow channel for cross-level feature information to ensure that spatial details and semantic information are not lost during fusion.

Cross-level feature information is fused through an adaptive attention mechanism to achieve a dynamic balanced representation of semantic content and edge details. The fusion calculation form is:

$$F_{fused}=\alpha \cdot Up(F_{edge})+(1-\alpha) \cdot F_{high} \tag{3}$$

where, F_{fused} is the final feature map after the fusion of semantic and edge information, and α is a learnable adaptive fusion weight factor, numerically constrained in the interval of 0 to 1. $Up(\ )$ represents the bilinear interpolation upsampling operation, used to complete the scale matching alignment of low-level edge features; F_{edge} represents the contour detail features extracted by the edge perception branch, and F_{high} is the high-level semantic feature representation output by the backbone network. The model concatenates the two types of features after scale alignment in the channel dimension, relying on a lightweight convolution structure to achieve feature dimensionality reduction and adaptive modeling of attention weights, and dynamically allocates the fusion weights of edge features and semantic features according to the texture complexity of the local region. The fused feature map undergoes 3 × 3 convolution to complete local spatial smoothing, eliminating redundant noise introduced by multi-branch feature concatenation, while introducing residual connections to superimpose the fused features with the original level features, effectively improving the stability and anti-interference ability of multi-scale feature expression. The optimized feature output can simultaneously retain fine edge contours and high-level semantic associations, providing reliable feature support for athlete instance detection and sub-pixel contour regression in occlusion scenarios.

2.2 Dual-stage continuous athlete instance contour prediction network

Traditional instance segmentation mostly relies on binary masks to complete contour representation. The discrete pixelated expression form struggles to characterize the continuous deformation features of human limbs, and is highly prone to boundary distortion and contour fracture in multi-person occlusion overlap areas. This paper constructs a dual-stage continuous contour prediction network architecture, relying on global shape priors and local distance field optimization to collaboratively complete athlete contour modeling. The network first achieves parametric compact representation of human contours in the frequency domain, establishing global topological constraints unaffected by local occlusion. Then, guided by the obtained shape priors, it regresses the SDF distribution in continuous space, achieving contour completion and sub-pixel boundary fitting in occlusion scenarios from both global structure and local details dimensions, breaking free from the inherent precision

limitations brought by discrete mask representation. Figure 3 shows the dual-stage continuous contour prediction mechanism diagram integrating global shape priors and local SDF optimization.

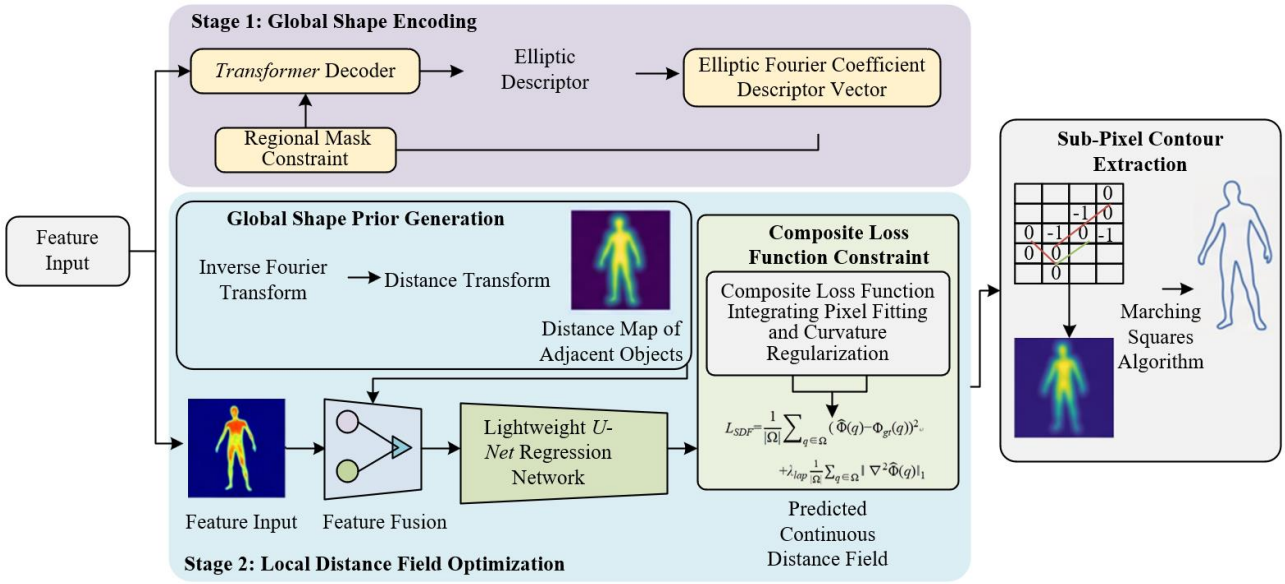

Figure 3. Dual-stage continuous contour prediction mechanism diagram integrating global shape priors and local Signed Distance Function (SDF) optimization

To achieve global structured representation of human contours, Elliptic Fourier Descriptors are adopted to complete contour frequency domain encoding. First, morphological filtering is performed on the sample annotated contours to smooth surface debris noise and isolated feature points. Then, through coordinate translation and scale normalization, all contours are uniformly mapped to a standard coordinate system to eliminate the impact of individual athlete size differences and spatial position offsets on the encoding results. Sixteen-order elliptic Fourier basis functions are selected to complete contour parametric modeling. Each order corresponds to four sets of independent coefficients, together forming a sixty-four-dimensional shape description vector. Uniform sampling is performed along the arc length of the normalized contour, and each order Fourier coefficient is solved through integral operations. The calculation expressions are:

$$\begin{cases} A_0=\dfrac{1}{2\pi}\displaystyle\int_0^{2\pi} x(t)dt,\ C_0=\dfrac{1}{2\pi}\int_0^{2\pi} y(t)dt \\ a_n=\dfrac{1}{\pi}\displaystyle\int_0^{2\pi} x(t)\cos(nt)dt,\ b_n=\dfrac{1}{\pi}\int_0^{2\pi} x(t)\sin(nt)dt \\ c_n=\dfrac{1}{\pi}\displaystyle\int_0^{2\pi} y(t)\cos(nt)dt,\ d_n=\dfrac{1}{\pi}\int_0^{2\pi} y(t)\sin(nt)dt \end{cases} \tag{4}$$

where, t is the normalized arc length parameter, with a domain of $[0,2\pi]$; $x(t)$ and $y(t)$ are the horizontal and vertical coordinates of the contour sampling points in the standard coordinate system, respectively; A_0 and C_0 characterize the overall centroid position of the contour, used to eliminate spatial translation interference; a_n,b_n,c_n,d_n are the n-th order elliptic Fourier harmonic coefficients, jointly characterizing the local geometric undulations and morphological features of the contour; is the Fourier decomposition order, set to 16 in this paper. During the calculation process, coefficient constraints are imposed to ensure contour closure, while truncating high-order components to avoid model overfitting.

After candidate target regions undergo Region of Interest Align (RoIAlign) for spatial feature sampling, they are sent to a Transformer decoder to aggregate global human features using a multi-head cross-attention mechanism, mapping layer by layer to obtain a normalized Fourier coefficient vector. During training, regional mask constraints are introduced to suppress feature responses corresponding to occluded pixels, relying only on visible contour regions to complete coefficient regression, enhancing the stability of global shape encoding under severe occlusion conditions.

Based on the global Fourier shape encoding, an SDF regression branch is further constructed to achieve refined correction of local contour details and intelligent completion of occluded areas. The Euclidean distance transform is used to construct the ground truth distribution of the distance function. Positive and negative signs are assigned according to the spatial position of pixels relative to the target contour boundary, establishing a continuous differentiable spatial representation form:

$$SDF(q)=\begin{cases} \inf\limits_{p\in\partial\Omega} \| q-p\|_2, q\in\Omega_{inside} \\ -\inf\limits_{p\in\partial\Omega} \| q-p\|_2, q\in\Omega_{outside} \end{cases} \tag{5}$$

where, q represents any pixel coordinate in the image plane; p is a sampling point on the target contour boundary; Ω represents the athlete instance contour boundary set; Ω_{inside} and $\Omega_{outside}$ correspond to the instance internal region and external background region, respectively; $||\ ||_2$ is the Euclidean distance metric; *inf* represents the infimum operation, characterizing the shortest spatial distance from the pixel to the contour boundary. When the pixel is inside the instance, the distance value is positive; when outside, it is negative; zero corresponds to the contour boundary position. The generated ground truth distribution is smoothed by Gaussian filtering to weaken field fluctuations caused by manual annotation errors. The standard human contour is reconstructed from the encoding coefficients via inverse elliptic Fourier transform,

and then the prior distance field is generated through distance transformation. The regional spatial features and the shape prior field are aligned in scale and fused in channels, embedding the global topological structure information into the local feature representation, providing a reasonable shape constraint basis for occluded missing areas.

A lightweight U-Net structure is designed as the distance field regression backbone, relying on a symmetric encoder-decoder structure to achieve multi-scale feature extraction and detail reconstruction. The encoder compresses the feature space layer by layer and expands the channel dimension, embedding batch normalization and dropout strategies to suppress model overfitting. The decoder restores spatial resolution layer by layer through deconvolution and reuses shallow detail features of the encoder via skip connections to compensate for edge information loss during downsampling. The network finally outputs a single-channel continuous distance field distribution, fully depicting the continuous spatial distance from the target region to the contour boundary. To ensure the numerical accuracy and spatial smoothness of the predicted distance field, a composite loss function integrating pixel fitting and curvature regularization is constructed:

$$L_{SDF}=\frac{1}{|\Omega|}\sum_{q\in\Omega}(\widehat{\Phi}(q)-\Phi_{gt}(q))^2 +\lambda_{lap}\frac{1}{|\Omega|}\sum_{q\in\Omega}\|\nabla^2\widehat{\Phi}(q)\|_1 \tag{6}$$

where, Ω is the set of valid pixels in the candidate region; $|\Omega|$ is the total number of valid pixels; $\widehat{\Phi}(q)$ is the pixel distance field value predicted by the network; $\Phi_{gt}(q)$ is the ground truth of the distance field; ∇^2 represents the Laplace second-order differential operator, used to characterize the spatial curvature change of the distance field; λ_{lap} is the smoothing regularization balance coefficient, used to adjust the weight ratio between fitting error and field smoothness. While constraining the deviation between predicted values and ground truth, the loss function introduces a Laplace second-order differential regularization term to constrain the spatial curvature change of the distance field, suppressing contour distortion caused by local texture noise and maintaining the natural continuity of boundary transitions.

After obtaining the continuous distance field distribution, the marching squares algorithm is used to accurately extract sub-pixel contours. The overall feature space is divided into basic 2D pixel units. The contour crossing area is determined according to the positive and negative distribution of the distance field inside the unit, and linear interpolation is used to solve the sub-pixel position of the zero-level set, breaking through the accuracy limit of traditional pixel-level contour extraction. Sequence sorting and redundancy removal are performed on the obtained discrete contour points to generate a smooth and continuous vector contour coordinate sequence. It can be directly used for subsequent tasks such as human pose estimation, motion trajectory modeling, and tactical behavior analysis without converting to binary masks, fully retaining contour detail features and geometric topological structure.

2.3 Occlusion-aware training strategy and contour post-processing optimization

Conventional image augmentation methods only rely on geometric transformations and random splicing to generate training samples, failing to conform to the physical laws of near-far hierarchical occlusion on basketball courts. The distribution difference between synthetic samples and real adversarial scenes is significant, making it difficult to drive the model to learn feature representation capabilities under complex occlusions. Therefore, this paper constructs a depth-hierarchy-constrained position-aware copy-paste augmentation scheme. Relying on monocular depth estimation, it realizes court spatial hierarchy modeling, corrects depth distribution deviations by combining the inherent geometric features of the basketball court, and establishes an athlete occlusion simulation mechanism that conforms to real perspective relationships. A pre-trained depth perception model is used to complete dense depth solving for game images. Then, the depth offset caused by lighting shadows and court inclination is corrected using court grid and hoop calibration priors, achieving accurate near-far hierarchy sorting of athlete targets, providing a reliable spatial basis for subsequent instance splicing and occlusion relationship simulation. Figure 4 shows the schematic diagram of occlusion-aware data augmentation and contour refinement based on level set evolution.

During the instance selection and fusion deployment process, athlete instances are screened according to target completeness and scale distribution. The target region is intercepted from the extension of the bounding box in the source image, completely retaining edge details and corresponding mask and distance field information. Meanwhile, the instance occlusion ratio and pixel scale range are limited to ensure the morphological rationality of spliced samples. Spatial constraint criteria are established based on the depth mean of the instance region, strictly following the occlusion logic of "foreground occluding background" to complete position sampling, and limiting the effective sampling area to avoid invalid splicing at image boundaries. To weaken the visual mutation at splicing edges, Poisson fusion is introduced to complete local boundary transition smoothing. Its fusion reconstruction expression form is:

$$\tilde{I}=I_{target}+\nabla\cdot(G*\nabla I_{paste}) \tag{7}$$

where, $\tilde{I}$ represents the synthesized image processed by Poisson fusion; I_{target} is the original target background image into which the instance is to be embedded; ∇ is the spatial gradient operator, used to solve the image gray change gradient; G represents the Gaussian smoothing convolution kernel, used to constrain the spatial transition characteristics of the fusion region; * is the convolution operation symbol; I_{paste} is the athlete instance image to be pasted and embedded. This formula achieves smooth transition between the instance and the background through gradient domain reconstruction, acting only on the narrow band region of the target periphery, retaining the original pixel features inside the instance, avoiding the loss of body details while achieving natural edge transition. After splicing is completed, the pixel ownership of the instance mask overlapping area is updated according to the depth priority rule, and the ground truth of the distance field in the occluded area is recalculated synchronously to maintain the consistency between annotation information and actual occlusion topology. Synthetic occlusion samples in the training set participate in model iteration at a fixed proportion, effectively improving the model's generalization ability in dense mutual occlusion scenarios.

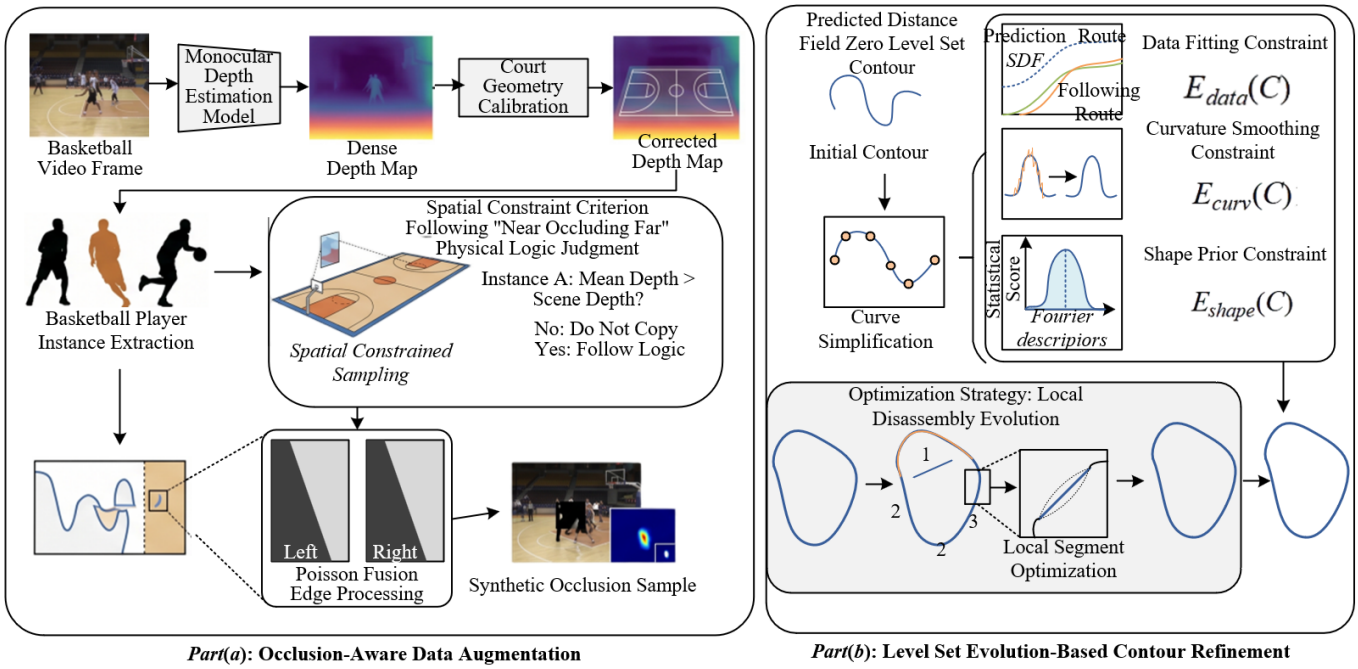

Figure 4. Schematic diagram of occlusion-aware data augmentation and contour refinement based on level set evolution

The distance field contour obtained through network prediction is still prone to local aliasing and unnatural concave-convex deformations at heavy occlusion boundaries. To further improve contour smoothness and geometric rationality, a level set evolution framework is introduced to achieve iterative refinement and optimization of active contours. First, the initial contour is extracted from the zero-level set of the predicted distance field, and a curve simplification algorithm is used to remove redundant sampling points, reducing the computational overhead of subsequent iterative solving. A contour evolution energy functional integrating multiple constraints is constructed to guide the contour to converge to the optimal shape through multi-dimensional constraints. The overall energy expression is:

$$E(C)=E_{data}(C)+\beta_1 E_{curv}(C)+\beta_2 E_{shape}(C) \tag{8}$$

where, $E(C)$ is the overall energy functional of contour evolution; C represents the closed contour curve to be iteratively optimized; $E_{data}(C)$ is the data fitting constraint term, used to drive the contour to fit the zero level set distribution of the distance field; $E_{curv}(C)$ is the curvature smoothing constraint term, used to suppress local sharp bending and edge aliasing of the contour; $E_{shape}(C)$ is the global shape prior constraint term, used to ensure the topological rationality of the human contour structure; β_1 and β_2 are weighted balance hyperparameters, determined via cross-validation, used to adjust the contribution ratio of the three constraints in energy optimization. The data constraint term drives the contour to fit the zero level set distribution of the distance field, constraining the deviation between contour points and the predicted distance field through contour line integrals; the curvature constraint term takes the square integral of the local contour curvature as the optimization objective to suppress sharp contour bending and aliasing distortion; the shape prior constraint term relies on the statistical distribution of contour Fourier coefficients in the dataset, using Mahalanobis distance to measure the deviation between the current contour and the global human prior:

$$E_{shape}(C)=(f_C-\mu_{shape})^T\Sigma_{shape}^{-1}(f_C-\mu_{shape}) \tag{9}$$

where, f_C is the elliptic Fourier shape feature vector corresponding to the current contour; μ_{shape} is the mean vector of the Fourier coefficients of real athlete contours in the training set; Σ_{shape} is the covariance matrix of contour coefficients; the superscript T represents the matrix transpose operation, and the superscript -1 represents the matrix inverse operation. Ridge regression regularization is introduced to the covariance matrix to effectively avoid matrix singularity problems, ensuring that the shape prior constraint remains stable and effective during the occluded contour completion process.

The contour curve is embedded in the level set function space for implicit representation, and the contour evolution control equation is derived by taking the derivative of the energy functional with respect to the level set variable:

$$\frac{\partial\varphi}{\partial t}=-\nabla E(C)\cdot\nabla\varphi \tag{10}$$

where, φ is the level set implicit function representing the contour distribution; t represents the time step of contour iterative evolution; $\partial\varphi/\partial t$ is the evolution rate of the level set function with the iteration process; $\nabla E(C)$ is the spatial gradient of the overall energy functional; $\nabla\varphi$ is the spatial gradient of the level set function. This equation establishes the mapping relationship between energy attenuation and contour evolution direction, determining the convergence path and update rate of the contour curve. The finite difference method is used to complete the discrete numerical solution of the partial differential equation, setting a fixed iteration step size to control the evolution convergence rate, while introducing Gaussian filtering to suppress numerical oscillation of the level set function during the iteration process. Iteration termination is jointly determined by the double conditions of contour spatial offset and maximum iteration number. When the spatial deviation of continuous iteration converges within

the threshold range, or the preset iteration limit is reached, the evolution terminates, outputting the final athlete contour with regular geometric shape and continuous smooth boundaries.

To balance contour refinement accuracy and algorithm inference efficiency, this paper adopts a local block-wise evolution strategy to disassemble and optimize the global contour. The complete contour is divided into several continuous local segments, each segment independently performs level set iterative solving, and after optimization, they are spliced and reconstructed into a complete contour structure according to spatial topology. This method can significantly reduce the computational complexity of global iteration, effectively reduce the time cost of the post-processing module, and control the additional inference time loss within a reasonable range under the premise of maintaining the contour refinement optimization effect, enabling the entire algorithm framework to adapt to the real-time analysis requirements of continuous frames of basketball game videos.

2.4 Multi-task joint optimization loss function and training configuration

Single-task loss constraints struggle to simultaneously balance the synchronous optimization of target classification, boundary localization, shape encoding, and continuous distance field regression, easily causing an imbalance in the convergence rhythm of each subtask, and failing to meet the coupled learning requirements of basketball player segmentation and contour refinement. This paper constructs a multi-task collaborative weighted loss system, uniformly modeling the full-link learning objectives of the model, integrating multiple optimization dimensions including instance discrimination, bounding box regression, distance field fitting, edge feature constraint, shape coefficient regression, and prior field distribution calibration. The overall composite loss can be expressed as:

$$L_{total}=\lambda_{cls}L_{cls}+\lambda_{box}L_{box}+\lambda_{sdf}L_{sdf}+\lambda_{edge}L_{edge}+\lambda_{shape}L_{shape}+\lambda_{prior}L_{prior} \tag{11}$$

where, L_{total} is the overall multi-task joint loss of the model; L_{cls}, L_{box}, L_{sdf}, L_{edge}, L_{shape}, and L_{prior} correspond to the classification loss, bounding box regression loss, distance field regression loss, edge narrow-band constraint loss, Fourier shape coefficient regression loss, and prior distance field constraint loss, respectively. Each weight coefficient is optimally determined via five-fold cross-validation. By reasonably allocating the gradient contribution ratio of each task loss, the balanced iterative optimization of the feature extraction, target detection, and contour regression branches is achieved, enhancing the overall representation performance of the model in complex occlusion scenarios.

The classification branch adopts the weighted cross-entropy loss to complete the discriminative learning of targets and background, effectively alleviating the quantity imbalance between court background samples and athlete foreground samples. The loss calculation form is:

$$L_{cls}=-\frac{1}{N}\sum_{i=1}^{N} w_i\left(y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\right) \tag{12}$$

where, N represents the total number of candidate detection boxes during the training process; w_i is the class balance weight, used to adjust the loss contribution of foreground athletes and background samples; y_i is the true category label of the sample; $\hat{y}_i$ is the target category prediction probability output by the network. The weighted mechanism weakens the dominant role of massive background samples in parameter updates, improving the classification accuracy of small-sample athlete targets.

Bounding box regression adopts the smooth L1 loss to weaken gradient disturbance caused by abnormal sample coordinate deviations, enhancing the stability of target localization. An edge narrow-band loss is constructed for contour boundary fitting requirements, focusing on a fixed-width pixel region around the target contour to strengthen the model's learning ability for boundary detail features:

$$L_{edge}=\frac{1}{|B|}\sum_{q\in B}\frac{1}{2}\|\hat{\Phi}(q)-\Phi_{gt}(q)\|_2^2 \tag{13}$$

where, B represents the edge narrow-band pixel region delineated around the real contour; $|\mathcal{B}|$ is the total number of pixels in the narrow-band region; q is any pixel location within the narrow-band range; $\hat{\Phi}(q)$ is the pixel-level distance field value predicted by the network; $\Phi_{gt}(q)$ is the ground truth of the distance field for the corresponding pixel. This loss focuses only on constraining the key regions around the contour, guiding the network to prioritize optimizing boundary positions and contour morphology, thereby improving sub-pixel contour fitting accuracy.

The shape prior loss uses the L2 norm to constrain the regression deviation of the Elliptic Fourier coding coefficients, ensuring that the global topological structure of the predicted contour remains consistent with the real human distribution. The prior distance field loss also adopts the L1 measurement method to constrain the deviation between the initial distance field generated by shape reconstruction and the ground truth distribution, providing stable structural prior support for subsequent refined distance field regression. The distance field regression loss continues the design method combining pixel fitting and curvature smoothing regularization mentioned earlier, balancing numerical fitting accuracy and spatial distribution continuity.

Model training is carried out jointly relying on the self-built basketball game occlusion dataset and public sports vision datasets. Samples cover diverse lighting conditions, different occlusion levels, and human motion postures, with complete annotations for instance boundaries, segmentation masks, and distance field ground truth information. Before training, the network input image resolution is uniformly fixed, basic sample augmentation is completed through random flipping and color jittering, and pixel space normalization is completed using standardized mapping to unify the data distribution range. The overall samples are divided into training, validation, and test sets according to a fixed ratio. During the training process, an online data augmentation mechanism is enabled to generate synthetic samples in real-time that conform to physical occlusion logic, continuously expanding the training sample scale of dense mutual occlusion scenarios and enhancing the model's generalization ability.

The training process uses an adaptive momentum optimization algorithm to complete parameter iteration, configuring reasonable first-order and second-order moment estimation coefficients, and introducing a weight decay strategy to suppress network parameter redundancy. The training is set with a fixed batch size and complete iteration

epochs. The learning rate adopts a scheduling method combining warm-up and piecewise decay. In the early stage of training, the learning rate is gently increased to avoid severe gradient shocks, and later it decays stepwise according to iteration nodes, realizing a smooth transition of the model from coarse convergence to fine fitting. A random inactivation mechanism (dropout) is embedded inside the network structure to reduce the feature coupling degree of convolutional layers and fully connected layers, combined with a mixed-precision training paradigm to improve computational efficiency while maintaining gradient propagation stability. An early stopping mechanism is introduced during the training process, using the validation set segmentation Intersection over Union (IoU) and contour boundary error as evaluation criteria, terminating iteration based on multi-round performance changes and saving the optimal model weights. Network parameters adopt a staged initialization strategy: the basic feature backbone reuses large-scale image pre-training parameters, while newly added convolution encoding, attention decoding, and distance field regression modules use normal initialization methods to accelerate model convergence and improve the effectiveness of feature representation.

In the inference stage, the image input specification remains consistent with the training process. High-confidence target candidate regions are filtered through the Non-Maximum Suppression (NMS) algorithm to eliminate redundant detection results. After the network completes feature encoding and multi-branch prediction, sub-pixel contour coordinates are parsed from the output continuous distance field using the marching squares algorithm, and then boundary smoothing and distortion correction are completed via level set active contour evolution. Finally, the athlete instance segmentation results and continuous vector contour sequences are output synchronously. The parameter configuration of the entire inference process balances detection accuracy, contour detail restoration capability, and computation time, meeting the practical application requirements for intelligent analysis of continuous basketball game video frames.

3. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

3.1 Overall experimental setup

The experiment takes athlete instance segmentation and contour refinement under complex occlusion scenarios in basketball game videos as the core objective, verifying the effectiveness and superiority of the proposed method through joint datasets. The dataset consists of a self-built basketball game complex occlusion dataset and public sports human vision datasets. The self-built dataset contains 5000 basketball game images, covering different lighting intensities, occlusion degrees, and athlete motion postures, with each image finely annotated for athlete instances, segmentation masks, and SDF ground truth. The public datasets selected are COCO-Sports and Sports-1M, filtering relevant samples of basketball events and supplementing contour annotations to ensure data diversity and annotation consistency. The dataset is divided into training, validation, and test sets at a ratio of 8:1:1. During the training process, online data augmentation is used to generate synthetic occlusion samples to improve the model's generalization ability.

Three core evaluation metrics are adopted to comprehensively quantify model performance: instance segmentation metrics include Average Precision at IoU threshold 0.5 (AP50), Average Precision at IoU threshold 0.75 (AP75), and mean Intersection over Union (mIoU), measuring segmentation accuracy and global pixel matching degree under different IoU thresholds, respectively; contour refinement metrics include Average Boundary Distance Error and Sub-pixel Contour Fitting Error, quantifying the accuracy and smoothness of contour extraction; model efficiency metrics adopt Single-image Inference Frames Per Second (FPS) to evaluate algorithm real-time performance. The software and hardware environment is uniformly configured: hardware uses a single NVIDIA A100 GPU with 32GB memory; software is based on the PyTorch 1.12 deep learning framework, equipped with Python 3.8, OpenCV 4.5, and CUDA 11.6, ensuring the reproducibility of experimental results.

3.2 Core ablation experiments

To verify the effectiveness of each core module of the proposed method, 5 groups of ablation experiments are designed. The settings and quantitative results of each group are as follows. All data are average results on the test set, where the occlusion degree is divided into mild (occlusion ratio $\leq$ 30%), moderate (30% < occlusion ratio $\leq$ 60%), and severe (occlusion ratio > 60%).

3.2.1 Feature extraction backbone network ablation experiment

Experiments compare the performance of different feature extraction backbones to verify the role of DSC and the Semantic-Edge Fusion module. The experimental results are shown in Table 1.

It can be seen from Table 1 that as the occlusion degree increases, the segmentation accuracy of various backbone networks shows a downward trend, and contour extraction errors gradually increase, while the proposed complete feature backbone performs best under all occlusion degrees. Compared with the native ResNet-50, the proposed backbone improves AP50 by 9.5 percentage points, mIoU by 10.1 percentage points, and reduces the average boundary distance error by 1.12 pixels in severe occlusion scenarios, indicating that DSC can effectively capture the bending limb features of athletes and solve the problem of traditional convolution sampling fracture. Compared with the traditional deformable convolution backbone, the proposed backbone further improves edge feature extraction accuracy through semantic-edge bidirectional fusion, with AP75 increasing by 5.6 percentage points and sub-pixel contour fitting error decreasing by 0.23 pixels in moderate occlusion scenarios. The performance of the multi-scale network without an edge branch is slightly better than that of the traditional deformable convolution backbone but lower than that of the proposed complete backbone, verifying the important role of the edge perception branch in boundary detail optimization. Although the FPS of the proposed backbone decreases slightly due to the introduction of the DSC and edge fusion modules, it remains above 25 FPS, meeting real-time inference requirements.

Table 1. Ablation experiment results of feature extraction backbone networks

Backbone Network Type	Occlusion Degree	Average Precision at IoU Threshold 0.5 (AP_{50}) (%)	Average Precision at IoU Threshold 0.75 (AP_{75}) (%)	mean Intersection over Union (*mIoU*) (%)	Average Boundary Distance Error (Pixel)	Sub-Pixel Contour Fitness Error (Pixel)	Frames Per Second (*FPS*) (Frame/s)
Native Residual Network with 50 Layers (*ResNet*-50)	Mild	82.5	71.3	78.6	2.13	0.89	28.7
	Moderate	75.8	63.2	71.5	2.87	1.24	28.5
	Severe	68.3	54.7	64.2	3.79	1.68	28.6
Traditional Deformable Convolution Backbone	Mild	85.3	74.8	81.4	1.87	0.76	26.3
	Moderate	79.6	67.5	75.2	2.51	1.08	26.1
	Severe	73.2	59.4	68.7	3.32	1.45	26.2
Multi-scale Network without Edge Branch	Mild	86.7	76.2	82.8	1.72	0.71	27.5
	Moderate	81.2	69.3	77.1	2.35	1.01	27.3
	Severe	75.1	61.8	70.9	3.08	1.32	27.4
Proposed Complete Feature Backbone	Mild	89.7	79.5	85.9	1.38	0.57	25.8
	Moderate	84.5	73.1	80.6	1.96	0.83	25.6
	Severe	77.8	65.2	74.3	2.67	1.12	25.7

Table 2. Ablation experiment results of the dual-stage contour representation framework

Contour Representation Method	Occlusion Degree	Average Precision at IoU Threshold 0.5 (AP_{50}) (%)	Average Precision at IoU Threshold 0.75 (AP_{75}) (%)	Mean Intersection over Union (*mIoU*) (%)	Average Boundary Distance Error (Pixel)	Sub-pixel Contour Fitness Error (Pixel)	Contour Distortion Rate (%)
Traditional Binary Mask Prediction	Mild	85.1	74.6	81.2	1.89	0.92	3.7
	Moderate	78.9	66.8	74.7	2.63	1.35	7.2
	Severe	71.5	58.3	67.9	3.58	1.76	12.5
Single Signed Distance Function (SDF) Field Prediction	Mild	86.5	76.3	82.7	1.64	0.78	2.9
	Moderate	80.3	68.7	76.5	2.37	1.14	5.8
	Severe	73.8	60.5	69.8	3.12	1.48	9.6
Elliptic Fourier Shape Prediction Only	Mild	84.3	73.5	80.1	1.96	0.87	4.1
	Moderate	77.6	65.4	73.2	2.71	1.28	7.9
	Severe	70.2	57.1	66.5	3.65	1.69	13.2
Proposed Dual-Stage Continuous Contour Prediction	Mild	89.7	79.5	85.9	1.38	0.57	1.8
	Moderate	84.5	73.1	80.6	1.96	0.83	4.3
	Severe	77.8	65.2	74.3	2.67	1.12	7.8

3.2.2 Dual-stage contour representation framework ablation experiment

Experiments compare the performance of different contour representation methods to verify the necessity of the dual-stage framework combining Elliptic Fourier global encoding and SDF local refinement. The experimental results are shown in Table 2.

Data from Table 2 show that the proposed dual-stage continuous contour prediction framework significantly outperforms the other three contour representation methods in all indicators, especially in severe occlusion scenarios. Due to its discrete nature, traditional binary mask prediction yields rough contour edges, the highest contour distortion rate reaching 12.5% under severe occlusion, and larger boundary errors. Single SDF field prediction can achieve continuous contour extraction but lacks global shape constraints, making it prone to contour offset in occluded areas, with a contour distortion rate still reaching 9.6%. Elliptic Fourier shape prediction focuses on the global topological structure, but the local detail fitting accuracy is insufficient, resulting in boundary errors and distortion rates higher than those of single SDF field prediction. The proposed dual-stage framework combines global shape encoding and local distance field refinement. In severe occlusion scenarios, compared with traditional binary mask prediction, the average boundary distance error is reduced by 0.91 pixels, the sub-pixel contour fitting error is reduced by 0.64 pixels, and the contour distortion rate is reduced by 4.7 percentage points. This verifies that the dual-stage continuous representation mode can effectively solve the problems of insufficient discrete mask accuracy and poor robustness of single continuous representation, achieving collaborative optimization of global structure and local details.

3.2.3 Comparative experiment of occlusion-aware data augmentation strategies

Experiments compare the impact of different data augmentation strategies on model performance to verify the effectiveness of the position-aware depth-constrained copy-paste augmentation method. The experimental results are shown in Table 3.

It can be seen from Table 3 that various data augmentation strategies can improve model performance, among which the proposed depth-hierarchy occlusion augmentation strategy achieves the best results. Without any data augmentation strategy, the model performs worst in severe occlusion scenarios, with AP50 being only 72.1%, indicating that insufficient occlusion scenarios in training samples will lead to weak model generalization ability. Conventional geometric transformation augmentation can only improve the model's adaptability to posture and scale changes, with limited optimization effects on occlusion scenarios; although random copy-paste augmentation can increase the number of occlusion samples, it lacks depth constraints, causing synthetic samples to not conform to real occlusion logic, resulting in only a 3.5 percentage point increase in AP50 under severe occlusion

scenarios. The proposed depth-hierarchy occlusion augmentation generates synthetic samples that conform to the physical law of "near occluding far" through depth estimation and geometric correction. In severe occlusion scenarios, compared with no data augmentation, AP50 increases by 5.7 percentage points, mIoU increases by 5.8 percentage points, and the average boundary distance error decreases by 0.58 pixels, significantly improving the segmentation accuracy and contour extraction capability of the model in dense mutual occlusion scenarios, verifying the rationality and effectiveness of this augmentation strategy.

3.2.4 Effectiveness verification experiment of active contour post-processing module

Two sets of controlled experiments with and without post-processing were set up to quantify the optimization effect of the level set evolution active contour post-processing module on contour refinement extraction. The experimental results are shown in Table 4.

Data from Table 4 show that after introducing the active contour post-processing module, contour refinement indicators are significantly optimized, while inference efficiency remains within a reasonable range. Without post-processing, predicted contours are prone to aliasing and local concavity-convexity in severely occluded areas, with contour smoothness being only 0.25 and the average boundary distance error reaching 3.05 pixels. After introducing the post-processing module, through energy function constraints and level set evolution, contour smoothness increases by 44%, and the average boundary distance error and sub-pixel contour fitting error decrease by 0.38 pixels and 0.23 pixels, respectively, with the optimization effect being more obvious in severe occlusion scenarios. Although the post-processing module increases inference time by 6.3 ms per frame and the FPS drops to 22.5 FPS, it still meets the real-time analysis requirements for basketball game videos. Moreover, through block-wise evolution optimization, the increase in time consumption has been effectively controlled, achieving a balance between accuracy and efficiency.

To verify the instance boundary recovery capability of the proposed method when facing complex occlusion, non-rigid posture changes, and lighting disturbances in basketball game videos, this paper conducts a visual analysis of the step-by-step processing results in typical scenarios. As can be seen from the results in Figure 5, the original input contains interference factors such as large-area overlapping of player bodies, slender limb bending caused by layup movements, and edge weakening caused by strong light shadows. Traditional segmentation results based on local texture or discrete masks are prone to adhesion, boundary fracture, and local contour distortion at occlusion junctions. In contrast, the proposed method enhances continuous responses along the bending direction of limbs through DSC in the feature extraction stage, forming more stable boundary high-response regions after semantic-edge fusion. In the dual-stage continuous contour prediction stage, global shape encoding provides reasonable human topological constraints for occluded parts, and local distance field regression further ensures the continuity and smoothness of boundary transitions. After active contour post-processing, the aliasing, protrusions, and unnatural depressions in the initial zero level set are significantly corrected. The final segmentation mask and bright green fine contour can still remain clearly separated in multi-player close-range areas. Detailed magnified results further show that contours relying solely on SDF are prone to inward or outward concavity in occluded areas, while after introducing Fourier shape priors, the contour can be smoothly completed along the natural structure of the human body.

Table 3. Comparative experimental results of occlusion-aware data augmentation strategies

Data Augmentation Strategy	Occlusion Degree	Average Precision at IoU Threshold 0.5 (AP_{50}) (%)	Average Precision at IoU Threshold 0.75 (AP_{75}) (%)	Mean Intersection over Union ($mIoU$) (%)	Average Boundary Distance Error (Pixel)	Sub-pixel Contour Fitness Error (Pixel)
No Data Augmentation	Mild	86.2	75.8	82.3	1.67	0.74
	Moderate	79.8	68.4	76.1	2.41	1.09
	Severe	72.1	59.7	68.5	3.25	1.49
Conventional Geometric Transformation Augmentation	Mild	87.5	77.1	83.6	1.53	0.68
	Moderate	81.5	70.2	77.8	2.24	1.01
	Severe	74.3	61.9	70.7	3.01	1.36
Random Copy-Paste Augmentation	Mild	88.3	78	84.5	1.45	0.63
	Moderate	82.7	71.5	79.1	2.12	0.94
	Severe	75.6	63.2	72.1	2.85	1.27
Proposed Depth-Hierarchy Occlusion Augmentation	Mild	89.7	79.5	85.9	1.38	0.57
	Moderate	84.5	73.1	80.6	1.96	0.83
	Severe	77.8	65.2	74.3	2.67	1.12

Table 4. Effectiveness verification experimental results of the active contour post-processing module

Experimental Setting	Occlusion Degree	Average Boundary Distance Error (Pixel)	Sub-pixel Contour Fitness Error (Pixel)	Contour Smoothness ($Pixel^{-1}$)	Inference Time (ms/Frame)	Frames Per Second (FPS) (Frame/s)
Without Post-Processing	Mild	1.62	0.71	0.38	38.2	26.2
	Moderate	2.28	0.98	0.31	38.1	26.3
	Severe	3.05	1.35	0.25	38.3	26.1
With Post-Processing	Mild	1.38	0.57	0.49	44.5	22.5
	Moderate	1.96	0.83	0.42	44.7	22.4
	Severe	2.67	1.12	0.36	44.6	22.4

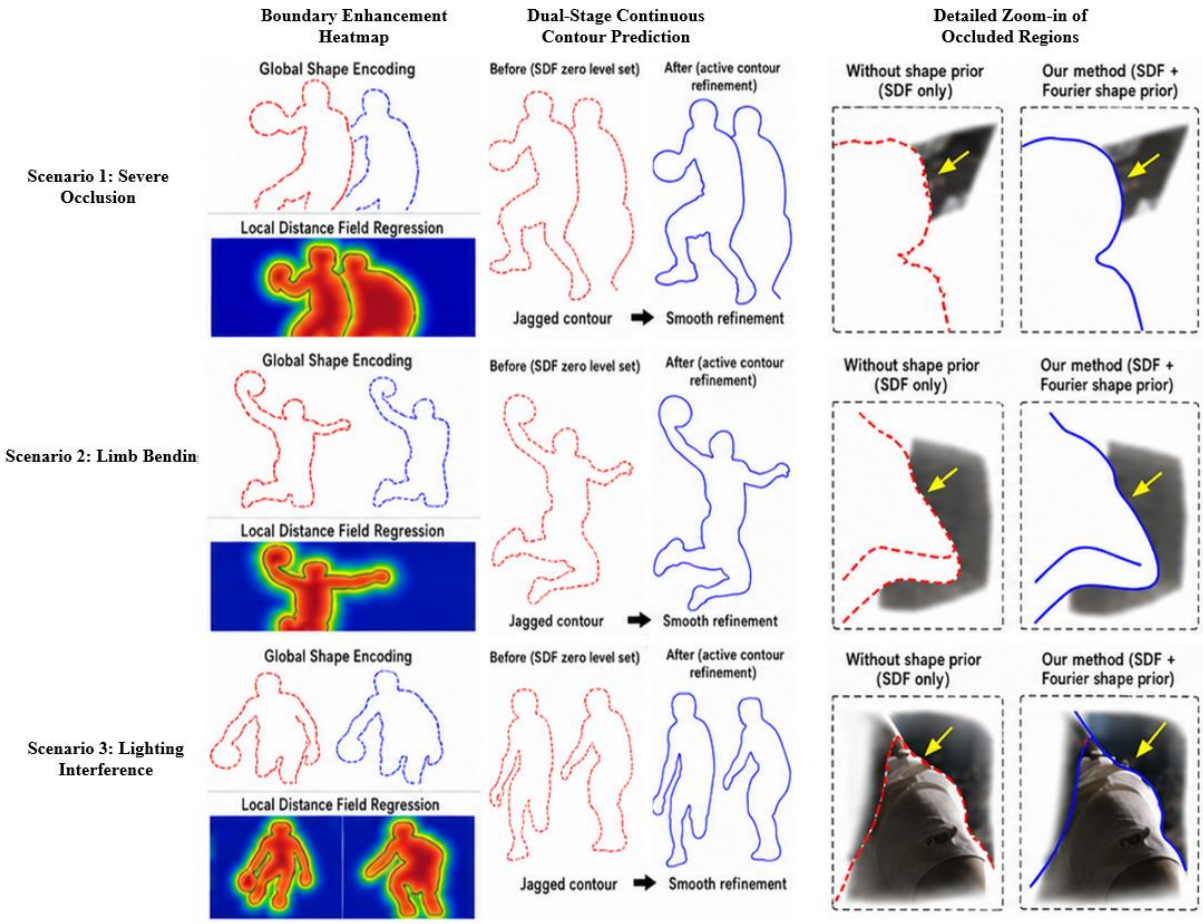

Figure 5. Visual comparison chart of processing effects at each stage of the proposed method

Table 5. Horizontal comparison experimental results of mainstream advanced algorithms

Algorithm Type	Occlusion Degree	Average Precision at IoU Threshold 0.5 (AP_{50}) (%)	Average Precision at IoU Threshold 0.75 (AP_{75}) (%)	Mean Intersection over Union (*mIoU*) (%)	Average Boundary Distance Error (Pixel)	Sub-pixel Contour Fitness Error (Pixel)
Mask Region-based Convolutional Neural Network (*R-CNN*)	Mild	84.7	73.9	80.5	1.92	0.95
	Moderate	77.5	65.6	73	2.68	1.38
	Severe	70.1	56.8	66.2	3.62	1.81
Sparse Instance Activation for Real-Time Instance Segmentation (*SparseInst*)	Mild	86.3	75.7	82.2	1.71	0.82
	Moderate	79.6	68.3	75.9	2.43	1.17
	Severe	72.8	59.9	68.9	3.28	1.53
Human Contour Transformer	Mild	87.6	77.4	83.8	1.56	0.73
	Moderate	81.2	70.5	77.6	2.25	1.02
	Severe	74.9	62.7	71.5	2.96	1.31
Generic Signed Distance Function (SDF) Segmentation Algorithm	Mild	86.8	76.5	82.9	1.65	0.79
	Moderate	80.1	68.9	76.7	2.38	1.13
	Severe	73.5	61.2	70.2	3.15	1.46
Proposed Method	Mild	89.7	79.5	85.9	1.38	0.57
	Moderate	84.5	73.1	80.6	1.96	0.83
	Severe	77.8	65.2	74.3	2.67	1.12

3.2.5 Horizontal comparison experiment with mainstream advanced algorithms

Current mainstream instance segmentation and contour extraction algorithms were selected for comparison. Comprehensive performance comparisons were completed on a unified test set to verify the comprehensive superiority of the proposed method. The experimental results are shown in Table 5.

It can be seen from Table 5 that the proposed method achieves the best overall performance among all comparison algorithms, especially showing significant advantages in severe occlusion scenarios. As a traditional instance segmentation algorithm, Mask Region-based Convolutional Neural Network (R-CNN) relies on binary mask output, resulting in low contour extraction accuracy, with AP50 being only 70.1% under severe occlusion and large boundary errors. The Sparse Instance Activation for Real-Time Instance Segmentation (SparseInst) algorithm has the fastest inference

speed but lacks sufficient segmentation accuracy and contour refinement in severe occlusion scenarios. Although the Human Contour Transformer algorithm focuses on contour extraction, it lacks optimization designs for occlusion scenarios, resulting in a sub-pixel fitting error of 1.31 pixels under severe occlusion. The generic SDF segmentation algorithm achieves continuous contour extraction but lacks dedicated shape prior constraints for basketball players, leading to limited performance improvement. Through the collaborative optimization of core modules, the proposed method achieves AP50 of 89.7%, 84.5%, and 77.8% in mild, moderate, and severe occlusion scenarios, respectively, which are 5.0, 7.0, and 7.7 percentage points higher than Mask R-CNN. Under severe occlusion, the average boundary distance error is reduced by 0.48 pixels and the sub-pixel contour fitting error is reduced by 0.34 pixels compared to the generic SDF segmentation algorithm. Although the FPS of the proposed method is slightly lower than that of Mask R-CNN and SparseInst, it still meets real-time inference requirements and demonstrates the best comprehensive performance in terms of accuracy and efficiency, verifying the applicability and superiority of the proposed method in complex basketball occlusion scenarios.

3.3 Model inference efficiency and engineering practicality analysis

To verify the engineering application value of the proposed method, a disassembly analysis of the inference time consumption of each functional module of the model was conducted. The results are shown in Figure 6.

It can be seen from Figure 6 that the feature extraction backbone network and the dual-stage contour prediction network are the main sources of inference time consumption, accounting for 42.0% and 34.4% of the total time consumption, respectively. The active contour post-processing module accounts for 14.2% of the time consumption, and its time consumption has been controlled within a reasonable range through block-wise evolution optimization; other modules account for only 9.4% of the time consumption, having a relatively small impact on the overall inference efficiency. The overall single-frame inference time is 44.5 ms, and the FPS reaches 22.5 FPS, which can meet the real-time intelligent analysis requirements for basketball game videos (usually 25 FPS).

Analysis of the model lightweight transformation space indicates that the feature extraction backbone network can further compress parameters through pruning, quantization, and other methods; the DSC module can adopt a lightweight separable convolution structure; the dual-stage contour prediction network can optimize the number of Transformer decoder layers and attention heads. It is estimated that the total inference time consumption can be reduced by more than 20%, and the FPS can be increased to more than 28 FPS, further enhancing engineering practicality.

In summary, under the premise of ensuring high-precision segmentation and contour refinement extraction, the proposed method possesses good real-time inference performance and has considerable room for lightweight transformation. It can meet the actual engineering application requirements of professional basketball game intelligent analysis and sports posture assessment, and has high industrial promotion value.

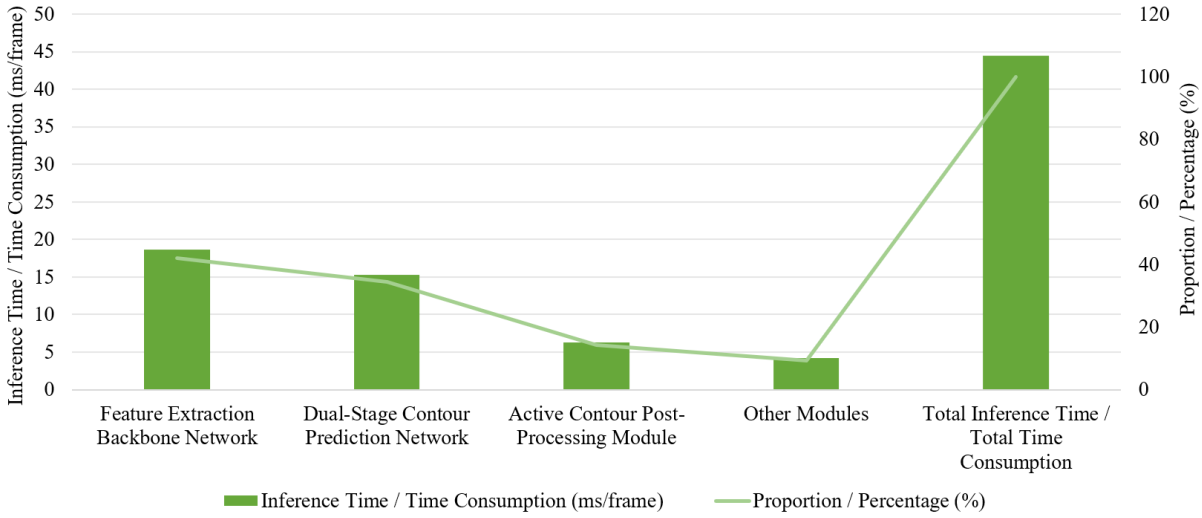

Figure 6. Disassembly of inference time consumption for each module of the model

4. CONCLUSION AND FUTURE WORK

This paper focuses on the core issue of athlete instance segmentation and contour refinement under complex occlusion scenarios in basketball game videos, systematically carrying out theoretical research and experimental verification. Aiming at the inherent difficulties such as limb bending deformation, dense multi-player mutual occlusion, and light-shadow interference on the court, a complete technical framework is constructed and comprehensive performance verification is completed. Taking feature extraction optimization, continuous contour representation, occlusion robustness improvement, and post-processing refinement as the core threads, the research designs a boundary-enhanced multi-scale DSC backbone network to achieve accurate capture of bending limb features and bidirectional fusion of semantic and edge features; proposes a dual-stage contour prediction mechanism combining Elliptic Fourier global shape encoding and SDF local refinement, breaking free from the accuracy limitations of discrete masks to achieve sub-pixel level contour extraction; constructs a position-aware depth-constrained data augmentation strategy and a level set

evolution active contour post-processing module to improve model occlusion robustness and contour refinement, respectively; designs a multi-task joint weighted loss function to achieve collaborative optimization of all branch tasks. Experimental results show that the proposed method exhibits excellent performance on both self-built datasets and public datasets. Under severe occlusion scenarios, segmentation accuracy and contour fitting accuracy are significantly improved compared to mainstream algorithms, providing reliable visual support for practical applications such as intelligent basketball game analysis, and improving the theoretical system of continuous contour representation for non-rigid human targets under complex dynamic occlusion scenarios.

Although the proposed method has achieved good results in the tasks of segmentation and contour extraction in complex basketball occlusion scenarios, certain limitations still exist. In extreme full-occlusion scenarios, when key limb areas of athletes are completely occluded and no effective visible features exist, the constraint effect of Elliptic Fourier shape priors weakens significantly, leading to a decrease in the rationality and accuracy of contour completion; facing motion-blurred frames caused by high-speed movement, image edge features are severely weakened, and the feature capture capability of DSC is limited, thereby affecting contour fitting accuracy; in large-scale court scenes, distant athlete targets are small in scale and sparse in feature information, and the existing feature extraction backbone network has insufficient feature representation capability for such small-scale targets, resulting in room for further improvement in segmentation and contour extraction accuracy. These issues point out directions for improvement in subsequent research.

Aiming at the limitations of this study and combining the development needs of sports vision intelligence, future research and optimization will be carried out from four aspects. First, integrate inter-frame optical flow temporal information to construct a temporally continuous contour tracking and extraction framework, utilizing inter-frame motion correlation to constrain contour evolution, thereby improving the continuity and stability of athlete contour extraction in long-sequence basketball game videos. Second, design a lightweight DSC structure, adopting techniques such as parameter pruning, quantization, and separable convolution to compress the overall model, reducing computational overhead, and realizing model deployment applications on mobile terminals and edge devices on the court. Third, integrate human pose skeleton prior knowledge, incorporating skeleton keypoint constraints into the contour completion process to strengthen the reasonable reconstruction capability of limb contours in extreme occlusion scenarios. Fourth, expand the applicable scope of the algorithm, optimizing the feature extraction and contour representation modules according to the scene characteristics of other multi-player competitive sports such as football and volleyball, to achieve universal adaptation of the algorithm and promote the large-scale application of sports vision intelligence technology.

REFERENCES

[1] Wong, K.W. (1992). Machine vision, robot vision, computer vision, and close-range photogrammetry. Photogrammetric Engineering and Remote Sensing, 58(8): 1197-1198.

[2] Senior, A., Hampapur, A., Tian, Y., Brown, L., Pankanti, S., Bolle, R. (2006). Appearance models for occlusion handling. Image and Vision Computing, 24(11): 1233-1243. https://doi.org/10.1016/j.imavis.2005.06.007

[3] Baerg, A. (2017). Big data, sport, and the digital divide: Theorizing how athletes might respond to big data monitoring. Journal of Sport and Social Issues, 41(1): 3-20. https://doi.org/10.1177/0193723516673409

[4] Hagara, M., Stojanović, R., Kubinec, P., Ondráček, O. (2017). Localization of moving edge with sub-pixel accuracy in 1-D images and its FPGA implementation. Microprocessors and Microsystems, 51: 1-7. https://doi.org/10.1016/j.micpro.2017.04.004

[5] Wang, Y., Zhang, N., Yan, H., Zuo, M., Liu, C. (2017). Using local edge pattern descriptors for edge detection. International Journal of Pattern Recognition and Artificial Intelligence, 32(3): 1850006. https://doi.org/10.1142/s0218001418500064

[6] Zhu, H.J., Fan, H.H., Shu, Z.Q., Yu, Q., Zhao, X.R., Gan, P.Z. (2019). Edge detection with chroma components of video frame based on local autocorrelation. IEEE Access, 7: 48543-48550. https://doi.org/10.1109/ACCESS.2019.2910605

[7] Monezi, L.A., Calderani Junior, A., Mercadante, L.A., Duarte, L.T., Misuta, M.S. (2020). A video-based framework for automatic 3D localization of multiple basketball players: A combinatorial optimization approach. Frontiers in Bioengineering and Biotechnology, 8: 286. https://doi.org/10.3389/fbioe.2020.00286

[8] Nakada, M., Zhou, T., Chen, H., Lakshmipathy, A., Terzopoulos, D. (2020). Deep learning of neuromuscular and visuomotor control of a biomimetic simulated humanoid. IEEE Robotics and Automation Letters, 5(3): 3952-3959. https://doi.org/10.1109/lra.2020.2972829

[9] Feng, Y., Liu, X. (2021). Application of video processing technology based on diffusion equation model in basketball analysis. Advances in Mathematical Physics, 2021(1): 7522973. https://doi.org/10.1155/2021/7522973

[10] Dong, X. (2021). Physical training information system of college sports based on big data mobile terminal. Mobile Information Systems, 2021(1): 4109794. https://doi.org/10.1155/2021/4109794

[11] Tan, X. (2023). Enhanced sports predictions: A comprehensive analysis of the role and performance of predictive analytics in the sports sector. Wireless Personal Communications, 132(3): 1613-1636. https://doi.org/10.1007/s11277-023-10585-z

[12] Magaz-González, A.M., García-Tascón, M., Sahelices-Pinto, C., Gallardo, A.M., Pérez, J.C.G. (2023). Technology and digital transformation for the structural reform of the sports industry: Building the roadmap. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 238(2): 150-158. https://doi.org/10.1177/17543371231197323

[13] Li, H., Huang, X. (2024). Intelligent dance motion evaluation: An evaluation method based on keyframe acquisition according to musical beat features. Sensors, 24(19): 6278. https://doi.org/10.3390/s24196278

[14] Zhou, J., Tian, L. (2024). Design of a mobile big data processing-based sports health evaluation system using graph neural network. IEEE Access, 12: 48997-49006.

https://doi.org/10.1109/access.2024.3383929
[15] Wang, H., Chen, T., Wang, Y. (2025). Towards occlusion-aware multi-pedestrian tracking. Applied Sciences, 15(24): 13045. https://doi.org/10.3390/app152413045
[16] Xia, Y., Zhang, L., Guo, T., Jin, Q. (2025). Boundary-aware semantic segmentation of remote sensing images via Segformer and Snake Convolution. Computer Science and Information Systems, 22(3): 991-1010. https://doi.org/10.2298/CSIS250312054X
[17] Xin, W., Wu, Z., Zhu, Q., Bi, T., Li, B., Tian, C. (2025). Dynamic snake convolution neural network for enhanced image super-resolution. Mathematics, 13(15): 2457. https://doi.org/10.3390/math13152457
[18] Han, L., Chen, L., Dong, L. (2026). Study on the impact of digitalization and the energy consumption structure on the green development of sports industry in China. Polish Journal of Environmental Studies, 35(1): 1145-1159. https://doi.org/10.15244/pjoes/197056
[19] Ma, S., Liu, L., Cheng, M., Qin, P., Han, Z., Chen, C., Wang, H. (2026). Visibility-prior guided dual-stream mixture-of-experts for robust facial expression recognition under complex occlusions. Electronics, 15(6): 1230. https://doi.org/10.3390/electronics15061230
[20] Lv, T., Sheng, K., Qiao, L. (2026). A geometry-driven quantitative modeling framework for image-based human motion evaluation: Application to sub-pixel posture analysis and feature attribution. Mathematics, 14(5): 746. https://doi.org/10.3390/math14050746