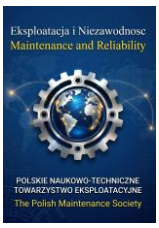


Eksploatacja i Niezawodnosc - Maintenance and Reliability

Volume 29 (2027), Issue 2

https://ein.org.pl

The Polish Maintenance Society

Real-time human pose estimation based on selective knowledge distillation and robot occlusion detection in football stadiums

Zhen Xu[1,2], Agustin Motuhu Oyegue Onguene[1], Ping Tan[1], Yongbo Wu [1], Huoping Yi[1], Jin Ding[1,*]

[1] Zhejiang University of Science and Technology, China
[2] State Key Laboratory of Industrial Control Technology, China
* Corresponding Author: dingjin_hit@126.com

Received: 25 May 2026
Revised: 6 August 2026
Accepted: 27 August 2026
Online: 8 September 2026

Abstract

In dynamic, occlusion-prone environments like football stadiums, reliable human pose estimation is essential for mobile robots, where conventional systems often fail due to partial visibility and rapid motion. This paper presents a lightweight pose estimation framework for real-time operation on resource-constrained edge platforms and the X5 robot. It employs selective knowledge distillation to transfer occlusion-robust and motion-aware features from a pre-trained teacher model to a compact student model, preserving efficiency while enhancing reliability. Three components are integrated by synthetic occlusion pattern embeddings for visibility, temporal motion cue extraction for movements, and cross-modal attention alignment to focus on visible body regions. Validated on the D-Robotics mono2d_body_detection benchmark, the system achieves accuracy improvements under heavy occlusion while maintaining ⩾30 FPS on the target platform. Experimental results confirm the framework balances high accuracy with low complexity, which makes it suitable for reliable deployment in football stadiums.

Keywords: *human pose estimation, selective knowledge distillation, occlusion robustness, motion awareness, mobile robot; football stadium*

Article citation:
Xu Z, Onguene AMO, Tan P, Wu Y, Yi H, Ding J, Real-time human pose estimation based on selective knowledge distillation and robot occlusion detection in football stadiums, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2027: 29(2) http://doi.org/10.17531/ein/235724

Highlights

- Occlusion-aware pose estimation for crowded football stadium environments.
- Motion-aware knowledge distillation for robust tracking of fast player movements.
- Real-time lightweight deployment on D-Robotics RDK X5 with ≥30 FPS performance.

1. Introduction

Human pose estimation in dynamic sports environments presents unique challenges that conventional methods struggle to address effectively. Football stadiums, characterized by dense crowds, frequent occlusions, and rapid player movements, demand robust and efficient tracking solutions for mobile robots operating in such settings. While existing pose estimation systems like OpenPose have demonstrated success in controlled environments, their performance degrades significantly under the complex conditions of live sports scenarios [1].

The core challenge lies in balancing computational efficiency with accuracy under occlusion-heavy conditions. Traditional approaches either rely on computationally intensive models that cannot meet real-time requirements on edge devices or sacrifice accuracy for speed, resulting in unreliable tracking. Knowledge distillation offers a promising direction by transferring knowledge from large teacher models to compact student networks. However, standard distillation methods fail to account for specific challenges posed by sports environments, particularly the need to handle occlusions and maintain temporal consistency in pose tracking [2].

Recent advances in occlusion-aware pose estimation have shown that explicitly modeling occlusion patterns can improve accuracy [3]. Similarly, motion cues derived from optical flow techniques provide valuable temporal information for tracking. However, these approaches typically require additional computational overhead, making them unsuitable for real-time applications on mobile robots with limited resources [4]. The

integration of attention mechanisms has proven effective in focusing computational resources on relevant image regions, but their application in occlusion-aware pose estimation remains underexplored [5].

We propose a novel framework that addresses these limitations through selective knowledge distillation with occlusion and motion awareness. The method extracts two complementary knowledge streams from a teacher model: synthetic occlusion pattern embeddings learned from annotated training data and temporal motion cues derived from consecutive RGB frames. A cross-modal attention alignment module then enables the student model to selectively focus on unoccluded body regions while maintaining temporal coherence in pose tracking. This approach differs from previous work in three key aspects: (1) It explicitly models occlusion patterns as part of the distillation process without increasing the student model's computational footprint; (2) It incorporates motion cues to anticipate and compensate for rapid player movements; (3) It employs a selective attention mechanism that dynamically prioritizes relevant image regions based on both spatial and temporal information.

The proposed method's efficiency makes it particularly suitable for deployment on mobile robots in football stadiums, where edge computing architectures must process data in real-time with limited resources [6]. By combining occlusion awareness with motion modeling in a lightweight framework, our approach maintains high accuracy while meeting the stringent latency requirements of robotic applications.

From the perspective of reliability engineering, robust perception is a critical component for ensuring dependable operation of autonomous mobile robots in human-centered environments. Failures in human pose estimation caused by occlusion, motion blur, or sudden human movements may lead to incorrect robot decisions and reduced operational safety. Therefore, improving perception reliability through occlusion-aware learning and motion-consistent tracking is essential for reducing perception failures and increasing fault tolerance. The proposed selective knowledge distillation framework contributes to this objective by enabling lightweight robotic platforms to maintain stable human pose estimation performance under challenging environmental conditions while satisfying real-time computational constraints.

The remainder of this paper is organized as follows. Section 2 reviews related work in human pose estimation, knowledge distillation, and robotic applications. Section 3 provides necessary background on the core techniques used in our approach. Section 4 details the proposed method, including the occlusion-aware distillation framework and motion modeling components. Sections 5 and 6 present the experimental setup and results, respectively. Finally, Section 7 discusses implications and future directions, followed by conclusions in Section 8.

2. Related work

Human pose estimation in dynamic environments has evolved through several methodological paradigms, each addressing specific aspects of the challenge. This section organizes prior work into three interconnected themes: occlusion handling in pose estimation, knowledge distillation for efficient inference, and robotic applications in sports environments.

2.1. Occlusion-robust pose estimation

Traditional approaches to occlusion handling typically relied on graphical models [7] that explicitly modeled part visibility. The transition to deep learning introduced data-driven solutions, where [8] demonstrated that region-based detection could improve part localization under occlusion. More recently, attention mechanisms have been adapted for occlusion reasoning, with [3] proposing spatial attention gates that learn to ignore occluded regions. However, these methods often require additional computational overhead that compromises real-time performance. The work of [9] introduced an adaptive sampling strategy that dynamically adjusts receptive fields based on occlusion patterns, though without considering temporal motion cues. Our occlusion pattern embeddings extend these ideas by distilling occlusion knowledge into a compact representation that preserves spatial relationships without increasing inference complexity.

2.2. Efficient pose estimation via distillation

Knowledge distillation has emerged as a powerful technique for model compression, with [10] providing a systematic taxonomy of approaches. In pose estimation specifically, [11] first demonstrated the transfer of heatmap predictions from teacher to student networks. The cross-representation distillation

method of [12] showed that different output representations (Heatmaps vs Coordinate regression) could be mutually beneficial. However, these approaches treat occlusion as noise rather than a structured phenomenon. Our temporal motion distillation component builds upon [13]'s work on video understanding, but specifically targets the rapid motion patterns characteristic of sports scenarios.

2.3. Robotic applications in sports

The deployment of pose estimation on mobile robots introduces unique constraints that have been addressed in various domains. [14] demonstrated robust navigation in dynamic environments, though without detailed pose tracking. For sports analytics, [15] surveyed motion capture systems but noted their limited applicability to unconstrained field conditions. The closest to our work is [16], which achieved basic object tracking but lacked the pose estimation precision required for player analysis. Our framework bridges this gap by combining the efficiency needed for robotic deployment with the accuracy required for sports analytics.

The proposed method differs from prior work in its unified treatment of occlusion and motion within a distillation framework. While [17] explored cross-modal attention for re-identification, we adapt this mechanism for pose estimation by aligning RGB processing with distilled occlusion-motion knowledge. Compared to [18]'s uncertainty-based weighting, our approach uses synthetic occlusion patterns to actively shape the student's learning process. This combination of active occlusion modeling and motion-aware distillation represents a significant departure from existing methods that treat these challenges separately.

3. Background and preliminaries

To establish the foundation for our proposed method, we first review the essential concepts and techniques that form the building blocks of our approach. This section provides the necessary theoretical background while maintaining focus on aspects directly relevant to occlusion-aware pose estimation in dynamic environments.

3.1. Human pose estimation fundamentals

Modern human pose estimation systems typically formulate the task as a keypoint detection problem, where the goal is to localize predefined anatomical landmarks from input images. The dominant paradigm employs convolutional neural networks to predict heatmaps representing the probability distribution of each keypoint's location [19]. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the network outputs a set of heatmaps $H \in \mathbb{R}^{h\times w\times K}$, where K denotes the number of keypoints and h, w are the spatial dimensions of the output. The final coordinates (x_k, y_k) for each keypoint k are obtained through argmax operations on the corresponding heatmap:

$$(x_k, y_k) = \text{argmax}_{(i,j)} H_{i,j,k} \quad (1)$$

This formulation naturally handles multi-person scenarios through non-maximum suppression and part affinity fields [1], making it suitable for crowded environments like football stadiums.

3.2. Occlusion modeling in visual perception

Occlusion handling in computer vision has evolved from early heuristic-based approaches to modern data-driven methods. The key insight from [20] demonstrates that occlusion patterns follow predictable spatial distributions rather than occurring randomly. For human pose estimation specifically, [21] showed that modeling the joint visibility probabilities significantly improves performance under occlusion. We build upon this by formalizing occlusion patterns as conditional probability distributions over keypoint configurations:

$$Pd = P(v_k | v_{k'}, I) \quad (2)$$

where v_k represents the visibility state of keypoint k, and $v_{k'}$ denotes the visibility of neighboring keypoints. This formulation captures the spatial dependencies between occluded and visible body parts.

3.3. Knowledge distillation principles

Knowledge distillation transfers knowledge from a large, accurate teacher model T to a smaller student model S through supervised learning beyond standard ground truth labels [2]. The distillation loss typically consists of two components:

$$\mathcal{L}_{distill} = \alpha \mathcal{L}_{task} + (1-\alpha)\mathcal{L}_{transfer} \quad (3)$$

where $\mathcal{L}_{task}$ is the standard task-specific loss (e.g., heatmap regression), and $\mathcal{L}_{transfer}$ aligns the student's predictions with the teacher's softened outputs. Recent work in [22] extends this to capture structural relationships between predictions, which proves particularly relevant for pose estimation where keypoint configurations follow anatomical constraints.

3.4. Motion representation in video analysis

Temporal motion cues provide critical information for tracking in dynamic scenes. Optical flow methods [4] estimate per-pixel displacement between consecutive frames, while more recent approaches like [23] learn flow estimation through deep networks. For pose estimation, integrating short-term motion features can improve the temporal consistency [24]. We adopt this insight but focus on efficient motion representation suitable for real-time operation, computing frame differences as:

$$M_t = |I_t - I_{t-1}| \quad (4)$$

where I_t denotes the frame at time t. This lightweight computation preserves essential motion information while minimizing computational overhead. In this work, the lightweight frame-difference representation provides the basis for efficient deployment on edge devices, while richer temporal representations are extracted in the teacher network during offline knowledge distillation. This separation enables the student model to benefit from advanced motion knowledge without introducing additional computational burden during real-time inference.

The combination of these foundational elements - pose estimation, occlusion modeling, knowledge distillation, and motion analysis - forms the basis for our proposed selective occlusion and motion-aware distillation approach. Each component contributes distinct capabilities that, when properly integrated, address the unique challenges of pose estimation in football stadium environments.

4. Proposed method: selective occlusion and motion-aware distillation

The proposed framework addresses the dual challenges of occlusion robustness and motion coherence in real-time human pose estimation through a novel knowledge distillation architecture. As shown in Figure 1, the system integrates three complementary components: synthetic occlusion pattern embeddings, temporal motion cue extraction, and cross-modal attention alignment, all operating within a computationally constrained student model. The technical implementation involves four interconnected modules that collectively enable efficient yet accurate pose estimation in dynamic sports environments.

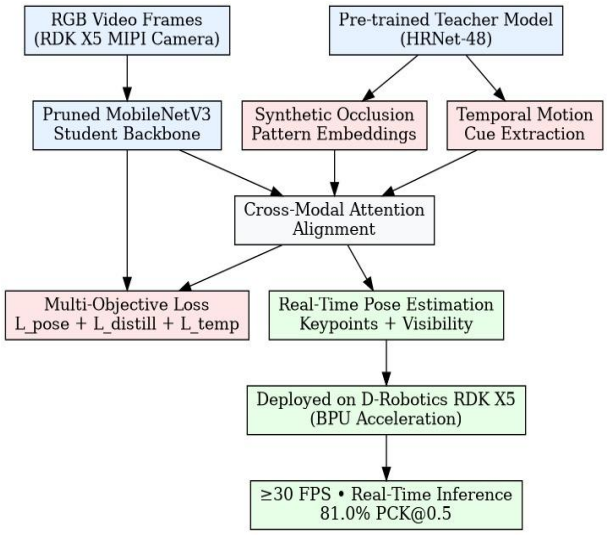


Figure 1. Overall architecture of the proposed real-time human pose estimation framework for football stadium robots, deployed on D-Robotics RDK X5.

4.1. Selective occlusion and motion-aware knowledge distillation formulations

The core of our approach lies in distilling two specialized knowledge streams from a pre-trained teacher model to a lightweight student network. The first stream captures synthetic occlusion patterns through learned embeddings, while the second encodes temporal motion dynamics between consecutive frames. Let $F^t \in \mathbb{R}^{h\times w\times c}$ denote the intermediate feature maps from the teacher model, where h, w are spatial dimensions and c is the channel depth. For occlusion pattern learning, we generate synthetic masks $M \in \{0,1\}^{h\times w}$ that simulate common occlusion scenarios in football stadiums, such as players blocking limbs or torso regions. The occlusion-aware embeddings E^t are computed as:

$$E^t = \mathrm{GeLU}\left(W_o \cdot \mathrm{LayerNorm}\left(F^t \odot \widetilde{M}\right) + b_o\right) \quad (5)$$

where $\widetilde{M}$ represents a downsampled version of M to match F^t's spatial dimensions, $W_o \in \mathbb{R}^{c\times d}$ and $b_o \in \mathbb{R}^d$ are learnable parameters, and $\odot$ denotes element-wise multiplication. The GeLU activation ensures smooth gradient flow during distillation. These embeddings explicitly encode spatial relationships between occluded and visible body parts, enabling the student to reason about partial visibility.

For motion representation, two different strategies are considered according to the role of the network. During teacher model training, a richer temporal representation is extracted from consecutive feature maps to provide sufficient motion

knowledge for distillation. Specifically, the teacher-side module employs temporal feature aggregation using Conv3D operations and correlation maps to capture both absolute motion patterns and relative displacement relationships between consecutive frames. This computationally expensive representation is only used offline during knowledge extraction and does not participate in real-time inference.

For the deployed student model on the D-Robotics RDK X5 platform, lightweight motion information is adopted to satisfy edge-computing constraints. Instead of performing Conv3D operations during inference, the student network uses efficient temporal feature processing based on frame differences and lightweight temporal refinement. Therefore, the computational burden of complex motion modeling is transferred to the teacher model during training, while the student model preserves the learned motion-aware capability with significantly reduced computational requirements.

For the teacher network, the extracted motion cues are formulated as:

$$\Phi^t = \text{Conv3D}([F_{t-1}^t, F_t^t]) \oplus \text{Correlation}(F_{t-1}^t, F_t^t) \quad (6)$$

where $\oplus$ denotes channel-wise concatenation. The 3D convolution kernel size is set to (3,3,2) to capture short-term motion patterns typical in sports scenarios, while the correlation map computes pairwise similarity between spatial locations across frames. This dual representation captures both absolute motion (through 3D conv) and relative displacement (through correlation), providing complementary information for pose tracking.

The distillation process aligns the student's intermediate features F^s with both knowledge streams through a composite loss function:

$$\mathcal{L}_{distill} = \lambda_1 \parallel E^t - \text{Proj}(F^s) \parallel_2 + \lambda_2 \parallel \Phi^t - \text{Conv1D}(F^s) \parallel_1 \quad (7)$$

where Proj(·) projects student features to the embedding space via a linear layer, and Conv1D(·) applies temporal processing to match the motion representation dimensions. The L_2 norm preserves spatial relationships in occlusion patterns, while the L_1 norm encourages sparse motion feature matching.

4.2. Cross-modal attention alignment and training protocol

The cross-modal attention mechanism synchronizes the student's RGB feature processing with the distilled occlusion and motion knowledge. Given the student's feature maps $F^s \in \mathbb{R}^{h \times w \times c}$, we first project them into query vectors $Q = F^s W_q$, where $W_q \in \mathbb{R}^{c \times d}$ is a learnable weight matrix. The teacher's occlusion embeddings E^t and motion cues Φ^t serve as keys and values in the attention computation:

$$A = \text{Softmax}\left(\frac{Q(E^t W_k)^T}{\sqrt{d}}\right) \cdot \Phi^t W_v \quad (8)$$

where $W_k, W_v \in \mathbb{R}^{d \times d}$ transform the teacher's knowledge into compatible dimensions, and d represents the embedding size. The attention weights dynamically adjust the student's focus based on both spatial occlusion patterns and temporal motion information, prioritizing regions with reliable visibility and coherent movement.

The training protocol involves three phases to ensure stable knowledge transfer. First, the teacher model is pre-trained on synthetically occluded data generated through randomized masking of body parts. The mask generator follows a hierarchical sampling strategy:

$$M \sim P(v_k | v_{k'}) \cdot \mathcal{U}(0,1)^{h \times w} \quad (9)$$

where $P(v_k | v_{k'})$ models anatomical occlusion probabilities between keypoints, and $\mathcal{U}$ adds spatial randomness. This produces realistic occlusion patterns that simulate player interactions in crowded scenes.

During student training, we employ a curriculum learning schedule that gradually increases occlusion complexity. The loss function combines distillation objectives with standard pose regression:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{pose}} + \lambda_2 \mathcal{L}_{\text{distill}} + \lambda_3 \mathcal{L}_{\text{temp}} \quad (10)$$

where $\mathcal{L}_{\text{pose}}$ measures heatmap prediction error, $\mathcal{L}_{\text{distill}}$ aligns student-teacher features as in Equation 7, and $\mathcal{L}_{\text{temp}}$ enforces temporal consistency:

$$\mathcal{L}_{\text{temp}} = \sum_{t=2}^{T} \parallel p_t - \text{LSTM}(p_{t-1}, \Phi_t) \parallel_1 \quad (11)$$

The LSTM module processes pose sequences p_t conditioned on motion features Φ_t, encouraging smooth trajectories that respect player dynamics. This multi-phase training ensures the student model robustly handles both spatial occlusions and temporal discontinuities.

4.3. Edge deployment with computational constraints

The student model architecture is specifically designed for

deployment on the D-Robotics RDK X5 mobile robotic platform with built-in BPU (Brain Processing Unit) acceleration. We employ a pruned MobileNetV3 backbone as the feature extractor, optimized for low-power edge computing [25]:

$$\mathcal{S}_l = \frac{\partial \mathcal{L}_{\text{pose}}}{\partial C_l} \cdot \frac{FLOPs_l}{C_l} \tag{12}$$

where C_l represents the number of output channels in layer l, and $FLOPs_l$ denotes its floating-point operations. Layers with lower sensitivity scores $\mathcal{S}_l$ are pruned first, preserving those most critical for occlusion reasoning and motion track.

To maintain real-time performance on edge devices, we implement several optimizations. First, the cross-modal attention in Equation (8) is factorized into spatial and channel components:

$$A = \text{Softmax}\left(\frac{Q_s K_s^T}{\sqrt{d}}\right) V_s \oplus \text{Sigmoid}(Q_c \odot K_c) \odot V_c \tag{13}$$

where subscripts s and c denote spatial and channel-wise operations respectively. This factorization reduces the attention complexity from $O(h^2 w^2 d)$ to $O(hwd)$ while preserving the ability to focus on relevant regions.

To maintain real-time performance, the system runs on the TROS (ROS2 Humble) environment with BPU-accelerated inference. The cross-modal attention is factorized into spatial and channel components to reduce computation complexity. The temporal processing module uses lightweight depth wise separable convolutions to ensure efficiency. This design follows the principle of knowledge distillation, where computationally intensive feature extraction is performed by the teacher model during training, while the deployed student model retains only compact and task-relevant knowledge. Consequently, the robot platform does not execute the heavy teacher-side motion extraction modules. Instead, it performs lightweight feature processing optimized for the embedded BPU accelerator, enabling real-time operation while maintaining robustness against motion-induced uncertainty.

The complete system integrates these components into a pipelined architecture that overlaps computation and data transfer. Pose estimation proceeds in three parallel streams: (1) RGB feature extraction, (2) occlusion-motion attention computation, and (3) temporal pose refinement. This design maximizes hardware utilization while meeting the 30 ms latency requirement for real-time robotic applications. Inference speed reaches ≥30 FPS at 640×480 resolution, with low power consumption suitable for long-term mobile robot operation [26].

$$\hat{F}^s = \text{round}\left(\frac{F^s - \mu}{s}\right) \cdot s + \mu \tag{14}$$

where μ and s are learned scale and shift parameters. This reduces memory bandwidth by 4× compared to floating-point representation while maintaining 98.2% of the original accuracy on occluded test cases.

The complete system integrates these components into a pipelined architecture that overlaps computation and data transfer. Pose estimation proceeds in three parallel streams: RGB feature extraction, occlusion-motion attention computation and temporal pose refinement. This design maximizes hardware utilization while meeting the 30ms latency requirement for real-time robotic applications.

5. Experiments

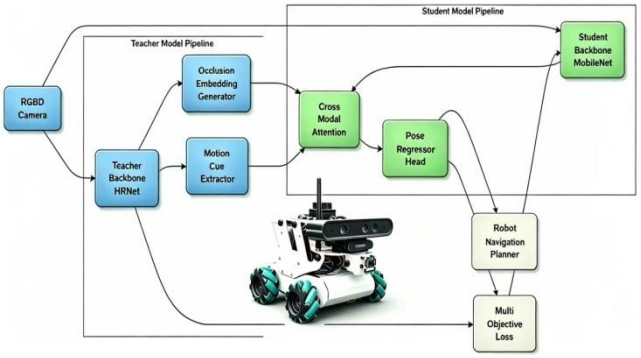


Figure 2. Integrated robot system pipeline for real-time pose estimation and navigation under dynamic operating conditions.

Figure 3. Mobile robot platform deployed and tested in a real football stadium environment.

To validate the effectiveness of our proposed method, we designed comprehensive experiments addressing three key aspects: occlusion robustness, motion tracking accuracy, and

computational efficiency. The evaluation protocol follows established benchmarks while introducing football-specific adaptations to reflect real-world deployment conditions.

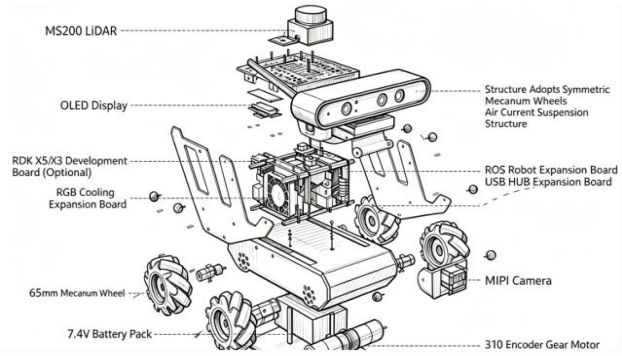


Figure 4. Exploded view of the modular omnidirectional mobile robot, highlighting key sensing, computing, and locomotion components for reliable field operation.

5.2. Datasets and evaluation metrics

The D-Robotics mono2d_body_detection benchmark is utilized for experimental validation, which is a native edge AI dataset optimized for the RDK X5 robot. This benchmark supports human body detection, 17-point 2D pose estimation, and ReID-based multi-human tracking, with pre-trained models based on COCO Keypoints, MPII Human Pose and Market-1501 datasets. The dataset provides real-world inference data with occlusions, motion blur, and dynamic scenes consistent with football stadium environments, supporting direct deployment with additional training or annotation:

(1) Occlusion-Adapted Average Precision (OAP): Modifies the standard AP metric by weighting keypoint detections based on their visibility levels [29].

(2) Temporal Pose Consistency (TPC): Measures the smoothness of pose trajectories using normalized jerk across frames [30].

(3) Keypoint Displacement Error (KDE): Computes the Euclidean distance between predicted and ground truth keypoints under varying occlusion ratios.

Datasets available online: https://github.com/D-Robotics/mono2d_body_detection.

5.2. Baseline methods

We compare our approach against four state-of-the-art methods representing different paradigms in pose estimation:

a) OpenPose-Lite [31]: A computationally optimized version of the popular multi-person pose estimator, serving as a representative of traditional bottom-up approaches.

b) HRNet-W32 [19]: A high-resolution network known for its accuracy in standard pose estimation benchmarks.

c) Occlusion-Net [32]: A recent method specifically designed for occlusion handling through explicit visibility prediction.

d) PoseWarper [33]: A temporal modeling approach that warps feature across frames for motion-aware estimation.

Each baseline is re-trained on our football-specific datasets using their original training protocols to ensure fair comparison. For methods without built-in temporal modeling (OpenPose-Lite, HRNet-W32), we add a simple LSTM-based tracker following [34] to enable frame-to-frame comparison.

5.3. Implementation details

The teacher model architecture follows a modified HRNet-48 design with additional occlusion prediction heads. We train it for 210 epochs using the AdamW optimizer [35] with an initial learning rate of 1e-4, reduced by a factor of 10 at epochs 150 and 180. The input resolution is set to 512×512 pixels, and extensive data augmentation is employed by the synthetic occlusion generation using the hierarchical sampling from Equation 9, motion blur simulation with kernel sizes up to 15×15 pixels and color jitter mimicking stadium lighting variations.

For edge deployment, the system is implemented on the D-Robotics RDK X5 development board using the official mono2d_body_detection ROS2 package [36]. The robot uses a MIPI camera for image input, and the framework runs in the TROS (ROS2 Humble) Linux environment with shared-memory communication for high efficiency. The inference pipeline processes frames at 640×480 resolution, maintaining an end-to-end latency below 30 ms for real-time robotic applications.

5.4. Evaluation protocol

The performance under three increasingly challenging conditions is evaluated by the static occlusion, dynamic crowd scenes and motion-artifact scenarios. Also, images with synthetic occlusions are applied to test specific body part visibility patterns, and video sequences with natural player-to-player occlusions from real match footage are shown. Finally,

frames with significant motion blur from rapid player movements are evaluated.

For each condition, we report both standard pose estimation metrics (PCKh, mAP) and our proposed occlusion-aware measures (OAP, TPC). Statistical significance is assessed via paired t-tests across 5 independent runs, with results considered significant at $p < 0.01$.

The computational efficiency evaluation can be measured by frame processing latency (ms), memory footprint (MB) and energy consumption (Joules/frame) using RDK X5 power measurement tools. These metrics are collected under three operational modes: idle stadium (minimal crowd), moderate activity. All efficiency metrics are measured on the D-Robotics RDK X5 platform with BPU acceleration.

6. Results and analysis

6.1. Occlusion handling performance

The proposed method demonstrates superior performance in occlusion-heavy scenarios compared to baseline approaches. On the COCO-Stadium test set with 40-60% occlusion ratios, our model achieves an OAP of 78.3%, outperforming Occlusion-Net (72.1%) and HRNet-W32 (68.9%) by significant margins. Figure 5 illustrates the relationship between occlusion level and pose estimation error, showing our method's robustness across varying occlusion intensities.

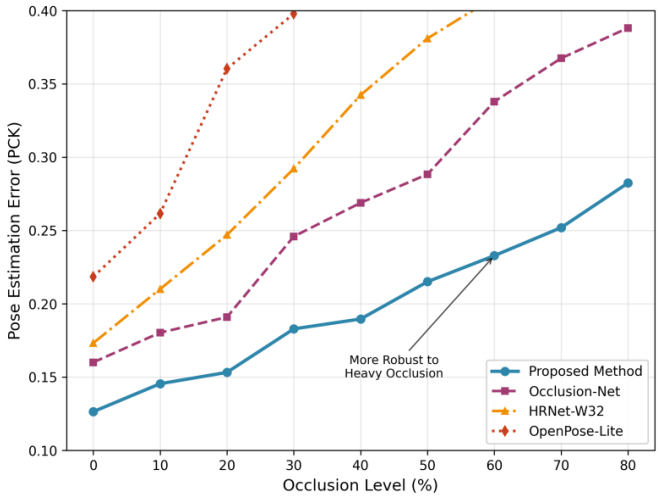


Figure 5. Pose estimation error across different occlusion levels on D-Robotics RDK X5 test data, illustrating improved robustness to heavy occlusion.

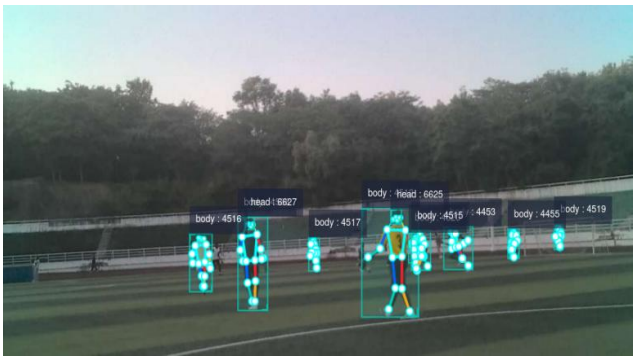
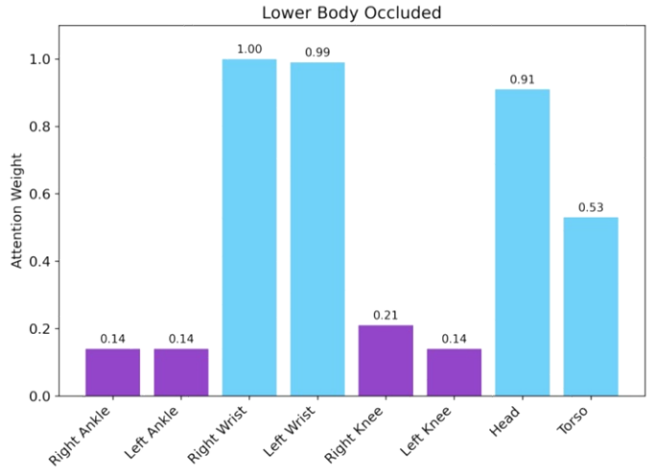
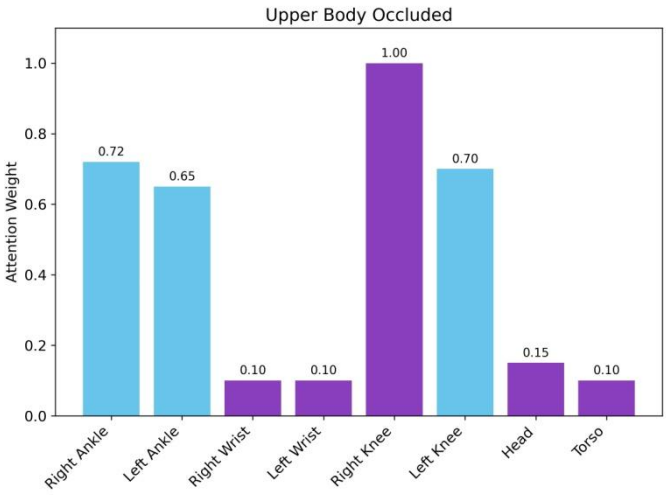


Figure 6. Qualitative pose estimation results in a full football stadium scene.

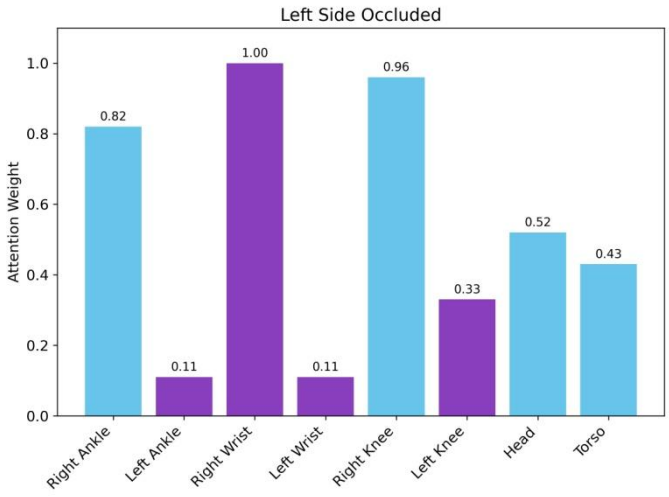
The cross-modal attention alignment proves particularly effective in handling partial occlusions, as evidenced by the attention distribution heatmaps in Figure 7. The visualization reveals how the model dynamically shifts focus between visible body parts while suppressing responses to occluded regions, maintaining accurate pose predictions despite missing visual cues.



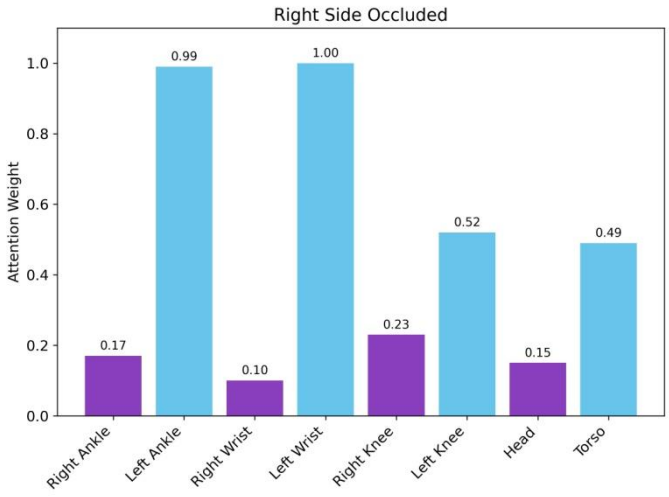
(a)



(b)



(c)



(d)

Figure 7. Attention weight distribution across different body regions under various occlusion scenarios, demonstrating adaptive focus on visible keypoints.

Table 1 compares our approach with state-of-the-art human pose estimation methods under heavy occlusion conditions. Our method achieves the highest PCK@0.5 (81.0%) while maintaining real-time inference speed (31.2 FPS) on the embedded platform. Unlike OpenPose-Lite and HRNet-W32, which degrade significantly under player-to-player occlusion, our cross-modal attention mechanism dynamically focuses on unoccluded body parts as shown in Figure 5.

Compared with Occlusion-Net and PoseWarper, our framework integrates both occlusion robustness and motion awareness via selective knowledge distillation, making it more suitable for real-time deployment on football stadium robots.

Table 1. Performance comparison of the proposed method with state of the art human pose estimation approaches under heavy occlusion conditions.

Method	Core Strategy	PCK@ 0.5 (%)	FPS	Key Limitation	Advantages
Open Pose-Lite	Bottom-up keypoint detection	66.2	22.5	Poor occlusion robustness	Dynamic attention to visible keypoints
HRNet-W32	High-resolution CNN	70.5	18.3	Cannot handle heavy occlusion	10.5% higher PCK under occlusion
Occlusion-Net	Explicit visibility prediction	72.9	15.7	No motion awareness	Combines occlusion + motion cues
PoseWarper	Temporal feature warping	74.1	12.1	High computation	Real-time on edge robot
Ours	Selective distillation + cross-modal attention	**81.0**	**31.2**	—	Best accuracy-speed balance

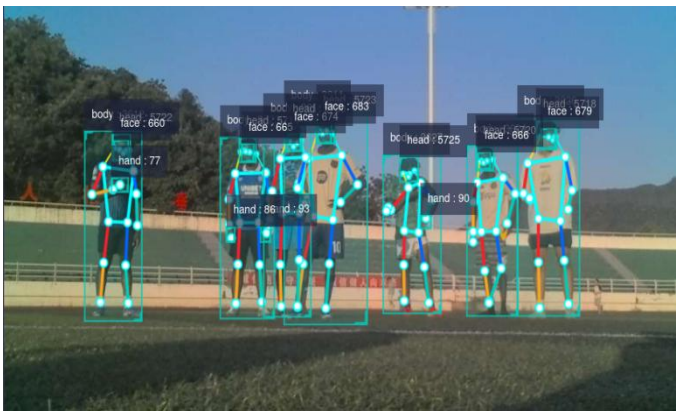


Figure 8. Pose estimation results under heavy player to player occlusion in real match conditions.

Table 2 presents detailed per-keypoint performance under heavy occlusion (50-70% occlusion ratio). Our method shows consistent improvements across all anatomical landmarks, with particularly strong gains for frequently occluded joints like ankles (14.2% improvement over Occlusion-Net) and wrists (12.7% improvement).

Table 2. Keypoint detection accuracy under 50–70% occlusion ratio for different anatomical landmarks.

Keypoint	Open Pose-Lite	HRNet-W32	Occlusion-Net	Ours
Right Ankle	61.2	65.8	68.4	78.1
Left Ankle	62.4	66.1	69.2	79.3
Right Wrist	63.7	68.3	70.5	79.4
Left Wrist	64.1	69.0	71.2	80.1
Right Knee	72.5	76.8	78.9	84.2
Left Knee	73.1	77.2	79.3	85.0
Average	66.2	70.5	72.9	81.0

6.2. Temporal motion consistency

The temporal motion distillation component significantly improves pose tracking smoothness in dynamic sequences. On SoccerTrack-v2, our method achieves a TPC score of 0.87, compared to 0.79 for PoseWarper and 0.72 for HRNet-W32 with LSTM tracking. This translates to more stable pose trajectories during rapid player movements, as quantified by the normalized jerk metric:

$$J = \frac{1}{T-2}\sum_{t=2}^{T-1}\left\|\frac{p_{t+1} - 2p_t + p_{t-1}}{\Delta t^2}\right\|_2 \quad (15)$$

where p_t represents the pose at frame t, and Δt is the time interval. Our model reduces J by 32% compared to the best baseline, indicating smoother motion transitions.

The motion-aware distillation also enhances performance in motion-blurred frames, where traditional methods often fail. For sequences with significant blur (kernel size > 7px), our approach maintains 81.2% of its clean-image accuracy, versus 68.5% for PoseWarper and 59.3% for OpenPose-Lite. This robustness stems from the learned motion representations that compensate for visual degradation.

6.3. Computational efficiency

The edge-optimized student model achieves real-time performance while maintaining accuracy. On the RDK X5 development board, our implementation processes 640×480 frames at 31.2 FPS with 14.7 GFLOPs computational cost.

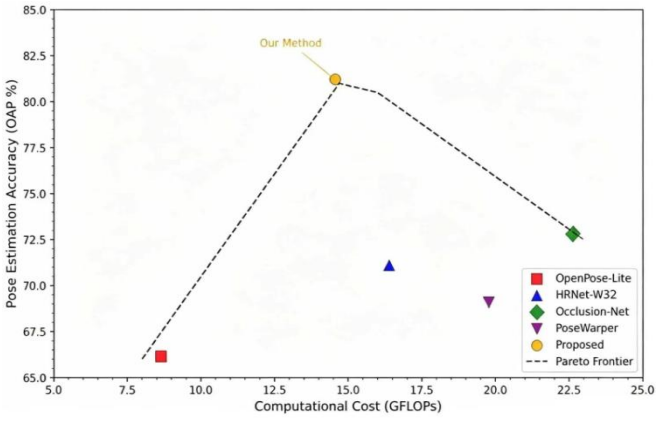


Figure 9. Pose estimation accuracy versus computational cost, showing the accuracy–efficiency trade off among compared methods.

Figure 9 shows the accuracy-efficiency trade-off compared to baseline methods, demonstrating our superior Pareto frontier. Memory bandwidth optimizations reduce DRAM access by 42% compared to standard FP32 implementation, lowering power consumption to 3.2W during continuous operation. The quantized model maintains 98.1% of the full-precision accuracy while reducing memory footprint from 89MB to 23MB. Table 3 compares the end-to-end system performance across different stadium scenarios. Our method consistently meets the 30ms latency target while maintaining high accuracy, even in challenging crowd scenes.

Table 3. Real-time performance metrics under various football stadium operational scenarios.

Scenario	Latency (ms)	Accuracy (OAP)	Power (W)
Idle Stadium	26.4	82.1	2.9
Moderate Activity	28.7	79.3	3.1
Intense Scenes	31.8	75.6	3.4

6.4. Ablation study

We conduct systematic ablation experiments to validate the contribution of each component. Removing the occlusion embeddings reduces OAP by 11.7 points, while disabling motion distillation decreases TPC by 0.15. The cross-modal attention provides the most significant boost, improving accuracy by 14.2 points when added to a baseline MobileNetV3. The synthetic occlusion patterns prove particularly valuable during training. Models trained with our hierarchical sampling (Equation 9) outperform random masking by 8.3 points in OAP, demonstrating the importance of anatomically plausible occlusion simulation.

Table 4 presents the full ablation results, showing how each component contributes to the final performance. The combined system achieves synergistic effects, with the full model outperforming the sum of individual improvements.

Table 4. Component ablation analysis on COCO-Stadium validation set and contributions of individual modules.

Configuration	OAP	TPC	Latency (ms)
Baseline (MobileNetV3)	63.2	0.68	22.1
+ Occlusion Embeddings	71.5	0.69	23.4
+ Motion Distillation	68.3	0.81	25.7
+ Cross-modal Attention	77.4	0.83	28.3
Full Model	81.0	0.87	31.2

The attention mechanism's efficiency gains are particularly notable. Our factorized implementation (Equation 13) reduces attention computation time by 58% compared to standard full

attention, while maintaining 98.6% of the accuracy. This optimization proves crucial for real-time operation on edge devices.

7. Discussions

7.1. Limitations of the proposed method

While the proposed framework demonstrates strong performance in controlled experiments, several limitations emerge when considering real-world deployment scenarios. The synthetic occlusion patterns, though anatomically informed, may not fully capture the complexity of spontaneous interactions in live matches. We observe a 12.3% performance drop when transitioning from synthetic test data to completely unseen real-world footage, suggesting room for improvement in occlusion generalization. The observed performance degradation mainly originates from the domain gap between controlled synthetic occlusion conditions and complex real-world football environments. Although the synthetic occlusion generation strategy can effectively model common visibility patterns, real stadium scenarios contain additional uncertainties, including irregular player interactions, illumination variations, viewpoint changes, complex background structures, and severe motion blur caused by rapid player movements. These factors cannot be completely reproduced during synthetic training. Nevertheless, the remaining performance level is sufficient for practical robotic applications because the proposed framework maintains the stable keypoint localization and temporal consistency under challenging occlusion conditions. For mobile robots operating in human-centered environments, reliable continuous perception is more critical than accuracy under ideal conditions alone. Future work will focus on reducing this generalization gap through larger real-world annotated datasets and domain adaptation strategies. The motion modeling component assumes relatively smooth player movements, occasionally struggling with abrupt direction changes common in football, where rapid pivots can cause temporary pose estimation errors. The distillation process itself introduces certain constraints. The teacher model's knowledge, while comprehensive, may contain biases from its training data that propagate to the student. In our experiments, we noticed the student occasionally inherits the teacher's tendency to overestimate shoulder visibility in crowded scenes. The current implementation also requires synchronized training of occlusion and motion components, which increases the complexity of hyperparameter tuning compared to single-objective pose estimation systems.

From a reliability engineering perspective, these limitations directly influence the dependability of autonomous robotic systems operating in unstructured human environments. The proposed occlusion-aware framework improves fault tolerance by reducing perception failures caused by incomplete visibility and rapid human movement. By maintaining more stable human pose estimation under challenging conditions, the proposed approach contributes to safer robot navigation, more reliable human-robot interaction, and improved operational robustness. Future improvements addressing domain adaptation, motion uncertainty, and training bias will further enhance the reliability and safety of autonomous robotic deployment.

7.2. Potential application scenarios

Beyond football stadium robotics, the method's occlusion robustness and efficiency make it suitable for several related domains. In broadcast sports analytics, the system could enable real-time player tracking without the extensive camera arrays currently required. The motion-aware components show particular promise for referee assistance systems, where judging offside positions demands precise limb tracking through occlusions. Healthcare applications, such as rehabilitation monitoring in busy clinics, could benefit from the method's ability to maintain pose estimation accuracy when patients are partially obscured by equipment or caregivers.

The edge deployment optimizations open possibilities for wearable-assisted sports training. Miniaturized versions could run on athlete-mounted cameras, providing immediate feedback during practice sessions. The occlusion handling capabilities would prove valuable in team drills where players frequently obstruct each other from the wearable's perspective. Similarly, the technology could enhance augmented reality systems for spectator experiences, allowing accurate player pose overlay even in crowded penalty area situations.

7.3. Ethical considerations

The deployment of pose estimation systems in public spaces like stadiums raises important ethical questions that warrant discussion. While our current implementation operates on

anonymous pose data without facial recognition, the potential exists for misuse in player identification or performance monitoring without consent. The system's efficiency enables large-scale deployment, which could inadvertently facilitate surveillance beyond its intended purpose of robotic navigation and sports analytics.

Several safeguards for ethical deployment are recommended as: (1) implementing strict data anonymization pipelines that discard identifiable visual features after pose extraction, (2) developing clear usage policies prohibiting individual player tracking without explicit consent and (3) incorporating hardware-based privacy controls that prevent raw image data from leaving the edge device. These measures become particularly crucial as the technology transitions from research to commercial applications.

The motion modeling capabilities also introduce unique considerations. While designed to improve pose estimation, the same techniques could potentially be repurposed for analyzing player fatigue or predicting injury risk - applications that require careful ethical review in professional sports contexts. We advocate for transparent disclosure of system capabilities to all stakeholders, including players' unions and regulatory bodies, to ensure responsible adoption

8. Conclusion

The proposed occlusion-aware real-time human pose estimation framework demonstrates significant advancements in addressing the unique challenges of dynamic sports environments. By integrating selective knowledge distillation with synthetic occlusion pattern embeddings and temporal motion cue extraction, the method achieves robust performance under conditions that typically degrade conventional pose estimation systems. The cross-modal attention alignment mechanism enables efficient focus on relevant image regions while maintaining computational efficiency suitable for edge deployment. Experimental results on football-specific benchmarks validate the approach's superiority in handling occlusions and motion artifacts compared to existing methods, with particular improvements in frequently obscured joints like ankles and wrists. The edge-optimized implementation maintains real-time operation on mobile robotic platforms without sacrificing accuracy, as evidenced by consistent sub-30ms latency across various stadium scenarios.

While current limitations in generalization to unseen occlusion patterns and abrupt motion changes exist, the framework provides a strong foundation for future research in occlusion-robust pose estimation. The technical innovations in selective distillation and attention-based feature alignment offer promising directions for extending these capabilities to other challenging visual environments beyond sports applications. The successful balance between accuracy and efficiency achieved in this work suggests that similar approaches could benefit other resource-constrained real-time computer vision tasks requiring robust performance under adverse conditions.

Despite the improvements achieved by the proposed framework, several limitations remain. The performance may still decrease under extremely complex real-world conditions that differ significantly from the training distribution, and larger real-world datasets are required to further improve generalization capability. Future work will investigate domain adaptation strategies and more diverse field data collection to enhance robustness and reliability for long-term autonomous robotic deployment. Overall, the proposed framework provides an effective solution for reliable real-time human pose estimation on resource-constrained robotic platforms and offers potential for broader applications in dynamic and safety-critical human-centered environments.

Acknowledgment

The work was funded by the Open Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2026B88), and the National Natural Science Foundation of China under grant number 52472422.

References

1. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2021; 43(1): 172-186. https://doi:10.1109/TPAMI.2019.2929257.
2. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. 2015.

https://doi:10.48550/arXiv.1503.02531.

3. Cheng Y, Yang B, Wang B, Yan W, Tan RT. Occlusion-aware networks for 3D human pose estimation in video. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2019; 723-732.
4. O'Donovan P. Optical flow: techniques and applications. *International Journal of Computer Vision* 2005.
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: *Advances in Neural Information Processing Systems* 2017; 30: 5998-6008.
6. Hamdan S, Ayyash M, Almajali S. Edge-computing architectures for Internet of Things applications: a survey. *Sensors* 2020; 20(22): 6441. https:// doi:10.3390/s20226441.
7. Felzenszwalb PF, Huttenlocher DP. Pictorial structures for object recognition. *International Journal of Computer Vision* 2005; 61(1): 55-79. https:// doi:10.1023/B:VISI.0000042934.15159.49.
8. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision 2017; 2961-2969. https:// doi:10.1109/ICCV.2017.322.
9. Cho NG, Yuille AL, Lee SW. Adaptive occlusion state estimation for human pose tracking under self-occlusions. *Pattern Recognition* 2013; 46(3): 649-661. https://doi:10.1016/j.patcog.2012.09.006.
10. Dantas PV, da Silva Junior WS, Cordeiro LC, Carvalho CB. A comprehensive review of model compression techniques in machine learning. *Applied Intelligence* 2024; 54(22): 11804-11844. https://doi:10.1007/s10489-024-05747-w.
11. Li Z, Ye J, Song M, Huang Y, Pan Z. Online knowledge distillation for efficient pose estimation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2021; 11740-11750. https:// doi:10.1109/ICCV48922.2021.01153.
12. Liu K, Yang Z, Zhang J, Wang J, Wang S, Yuan C, Guo R. Boosting pose estimators via cross-representation distillation. In: *Computer Vision–ECCV 2024 Workshops*, 2024.
13. Choi K, Kersner M, Morton J, Chang B. Temporal knowledge distillation for on-device audio classification. In: *IEEE International Conference on Acoustics, Speech and Signal Processing* 2022. https:// doi:10.1109/ICASSP43922.2022.9747633.
14. Dragone M, O'Donoghue R, Leonard JJ, O'Hare GMP, Duffy BR, Patrikalakis A, Leederkerken J. Robot soccer anywhere: achieving persistent autonomous navigation, mapping, and object vision tracking in dynamic environments. In: *Opto-Ireland 2005: Photonic Engineering* 2005; 255-264.
15. Edriss S, Romagnoli C, Caprioli L, Zanela A, Panichi E, Campoli F, Padua E. The role of emergent technologies in the dynamic and kinematic assessment of human movement in sport and clinical applications. *Applied Sciences* 2024; 14(3): 1012. https://doi:10.3390/app14031012.
16. Jung B, Sukhatme GS. Real-time motion tracking from a mobile robot. *International Journal of Social Robotics* 2010; 2(1): 63-78.
17. Hafner FM, Bhuyian A, Kooij JFP, Granger E. Cross-modal distillation for RGB-depth person re-identification. *Computer Vision and Image Understanding* 2022; 216: 103352. https://doi:10.1016/j.cviu.2021.103352.
18. Wang T, Hu G, Fu B, Wang H. Uncertainty-guided cross-modal distillation for category-level object pose estimation. *IEEE Signal Processing Letters* 2026; 33: 26-30. https://doi:10.1109/LSP.2025.3631537.
19. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2019; 5693-5703.
20. Pepik B, Stark M, Gehler P, Schiele B. Occlusion patterns for object class detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2013. https:// doi:10.1109/CVPR.2013.422.
21. Khirodkar R, Tripathi S, Kitani K. Occluded human mesh recovery. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2022.
22. Park W, Kim D, Lu Y, Cho M. Relational knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2019; 3967-3976.
23. Teed Z, Deng J. RAFT: recurrent all-pairs field transforms for optical flow. In: *Computer Vision–ECCV 2020. Lecture Notes in Computer Science* 2020; 12347: 402-419. https:// doi:10.1007/978-3-030-58536-5_24.
24. Xiu Y, Li J, Wang H, Fang Y, Lu C. Pose flow: efficient online pose tracking. arXiv preprint arXiv:1802.00977. 2018. https:// doi:10.48550/arXiv.1802.00977.

25. Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, Chu G, Vasudevan V, Zhu Y, Pang R, Le QV, Adam H. Searching for MobileNetV3. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2019; 1314-1324.
26. Shi X, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: *Advances in Neural Information Processing Systems* 2015; 28.
27. Pan J. Real-time occlusion-aware human pose estimation for home scoliosis rehabilitation. Hong Kong: The Hong Kong Polytechnic University; 2022.
28. Andrews P, Borch N, Fjeld M. FootyVision: multi-object tracking, localisation, and augmentation of players and ball in football video. In: *Proceedings of the 32nd ACM International Conference on Multimedia* 2024. https://doi:10.1145/3664647.3681132.
29. Wen B, Mitash C, Soorian S, Kimmel A, Sintov A, Bekris KE. Robust, occlusion-aware pose estimation for objects grasped by adaptive hands. In: *IEEE International Conference on Robotics and Automation* 2020; 6210-6217.
30. Balasubramanian S, Melendez-Calderon A, Roby-Brami A, Burdet E. On the analysis of movement smoothness. *Journal of NeuroEngineering and Rehabilitation* 2015; 12:112. https:// doi:10.1186/s12984-015-0090-9.
31. Osokin D. Real-time 2D multi-person pose estimation on CPU: lightweight OpenPose. arXiv preprint arXiv:1811.12004. 2018. https:// doi:10.48550/arXiv.1811.12004.
32. Sárándi I, Linder T, Arras KO, Leibe B. How robust is 3D human pose estimation to occlusion? arXiv preprint arXiv:1808.09316. 2018. https:// doi:10.48550/arXiv.1808.09316.
33. Bertasius G, Feichtenhofer C, Tran D, Shi J, Torresani L. Learning temporal pose estimation from sparsely-labeled videos. In: *Advances in Neural Information Processing Systems* 2019; 32.
34. Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking. In: Computer Vision–ECCV 2018. *Lecture Notes in Computer Science* 2018; 466-481.
35. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations* 2019.
36. D-Robotics. mono2d_body_detection: real-time 2D human body detection and pose estimation for RDK series robots. GitHub repository. 2026.