A Pose Estimation Pipeline for Sports Dive Analysis

Prabhu Nithin Gollapudi, Matrikelnummer: 23269243

Friedrich-Alexander-Universität Erlangen-Nürnberg
Machine Learning and Data Analytics Lab
https://www.mad.tf.fau.de/
Carl-Thiersch-Straße 2b, 91052 Erlangen, Germany
Email: prabhu.nithin.gollapudi@fau.de

***Abstract*—In modern sports, detailed motion analysis is the key to improving technique, preventing injuries, and supporting objective performance evaluation. As demands for precision and real-time feedback increase, computer vision techniques such as human pose estimation are becoming central to modern sports analytics. However, in dynamic disciplines such as diving, pose estimation systems must overcome challenges related to rapid motion, occlusion, complex body orientations, and real-time processing demands. This study presents a pose estimation pipeline designed specifically for diving analysis. It combines segmentation techniques to accurately isolate the diver from the background with pose estimation algorithms that track body keypoints and generate key performance indicators (KPIs) across different stages of the dive. These KPIs support training optimization, coaching feedback, and flight phase evaluation. Experimental evaluations on real-world diving footage demonstrate that our approach achieves a 200% improvement in frames processed per second (FPS) without sacrificing accuracy. This work can be extended towards real-time, AI-driven analytics in other broader sports performance contexts.**

I. Introduction

Today, the use of artificial intelligence (AI) and data-driven approaches to improve athletic performance has become quite common in sports. These technologies are transforming how athletes train, recover, and compete by providing insight into their performance. Among these technologies, human pose estimation stands out because it provides deep insight into the movements of an athlete [1] that allows coaches and trainers to optimize their training and refine techniques. Its applications span a wide range of sports, from running and gymnastics to swimming and diving, where precise biomechanical feedback can significantly influence outcomes. In fast-paced sports such as diving, where precise posture and body control are critical, pose estimation faces unique challenges, particularly due to rapid movements, complex rotations, and occlusions [2].

Traditional dive evaluation relies on human judges who assess various aspects of performance based on subjective criteria [3]. Although this approach has been the standard for decades, it is inherently influenced by factors such as bias, viewing angle, and momentary judgment. These inconsistencies can lead to variation in scoring, even for similar performances. Jakab et al. (2023) conducted an exploratory investigation of these traditional methods, highlighting their potential limitations and exploring opportunities for the development of AI-based systems [4]. Automated, data-driven approaches offer the potential to provide frame-level analysis and support more transparent feedback for athletes and coaches.

A more domain-specific approach, DiveNet [5], based on deep convolutional neural networks, focuses on the localization of dive actions and the extraction of physical parameters. Although DiveNet successfully detects the start and end of dive sequences and captures select motion attributes, it lacks continuous full-body pose tracking, KPI extraction, and comprehensive dive stage detection. This limits its ability to capture detailed biomechanical insights that are essential for performance evaluation. For example, Park and Yoon (2017) found that joint angles and angular velocities, particularly in the hip and knee, were key factors in distinguishing skilled from less skilled divers during back somersault pike dives [6]. Without continuous pose estimation, systems like DiveNet are unable to track these kinematic indicators throughout the full dive sequence. A system capable of full-body tracking and KPI extraction can bridge this gap by enabling fine-grained frame-level analysis across all phases of the dive.

Given the inherent limitations of traditional human-involved dive evaluation and the existing pose estimation methods, this study introduces a novel pose estimation pipeline specifically designed for comprehensive dive pose estimation and performance analysis. The proposed system enables continuous full-body tracking and detailed performance assessment in all phases of a dive, offering a more objective, data-driven alternative to current approaches.

II. Methods

A. Design

The proposed pipeline consists of three core components:

- Segmentation: A Robust Video Matting (RVM) [7] based approach that isolates the diver from the background.
- Pose Estimation: A Real-Time Multi-Person Pose Estimation (RTMPose) [8] approach with a real-time object detector (RTMDet) [9] + that provides continuous keypoint tracking for each frame.
- KPI generation: A framework for extracting key performance metrics, such as joint angles, height, and total rotation.

The structure of the system is illustrated in Figure 1.The system balances real-time performance with modularity, allowing each stage to be optimized independently. It also ensures that the system can be easily extended or upgraded without a complete redesign.

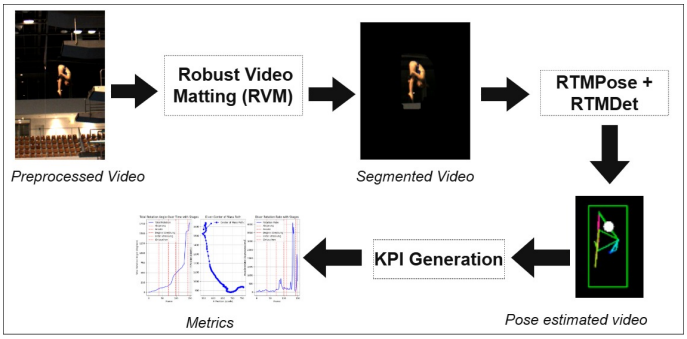

Fig. 1. Visual representation of the pipeline for diving analysis, illustrating the three main stages: segmentation, pose estimation, and KPI generation.

B. Data Collection

The video footage used in this study was provided by Olympiastützpunkt (OSP) Berlin[1]. The data set consists of many different recorded dives performed under training conditions by competitive athletes. The videos capture various phases of the dive with sufficient resolution (2048*1152 pixels) and frame rate (25 FPS) to support frame-level pose estimation and kinematic analysis. No personal identifiers were included and all data was used with permission for research purposes only.

C. Experimental Setup

All experiments were carried out on a local machine equipped with an NVIDIA RTX 3060 GPU, Intel Core i7-11800H CPU clocked at 2.3 GHz and 16 GB of RAM. The pose estimation pipeline was implemented in Python using version 3.12.2. All processing and performance evaluations, including FPS measurements and KPI extraction, were carried out on this system without the use of external servers or cloud computing resources.

D. Segmentation

For accurate pose estimation and kinematic analysis of divers, precise segmentation from the background is essential. The fast and complex movements inherent in diving, coupled with visual complexities such as water spray and reflections, make this a challenging task. We therefore evaluated the effectiveness of several existing segmentation methods: Segment Anything (SAM) [10], MediaPipe Selfie Segmentation [11], and YOLOv5 [12] offer good performance in general visual tasks, but are not optimized for the task at hand. In our tests, these methods struggled with precision and latency, making them less suitable for frame-by-frame analysis of diving footage.

To address these limitations, we adopted RVM, a recurrent neural network-based segmentation model designed for high-quality, low-latency background matting. RVM demonstrated superior performance in separating the diver from complex backgrounds across all frames, providing cleaner inputs for pose estimation while maintaining low computational overhead. Notably, RVM significantly reduced the file size from 366 MB to 23.2 MB (as detailed in Table I), achieving a decent processing speed of 3.46 FPS. This substantial compression offers considerable advantages in terms of storage efficiency and lowers computational demands for subsequent pose estimation. Its ability to handle temporal consistency across video frames made it particularly suitable for our application.

[1]We thank Michael Brunner for providing access to the diving video dataset used in this research.

TABLE I. COMPARISON OF SEGMENTATION TECHNIQUES ON INPUT VIDEO OF SIZE 366 MB AND 1328 FRAMES.

Technique	Total Time (s)	FPS	Output Size (MB)	Quality
RVM (mobilenetv3)	383.41	3.46	23.2	Excellent
SAM (vit_b)	701.99	1.89	163.5	Bad
Mediapipe	69.85	19.01	80.9	Good
YOLOv5	550.85	1.79	8.5	Good

E. Human Pose Estimation

RTMPose-m was selected for its efficient architecture and its recognition as a state-of-the-art (SOTA) model for real-time, multi-person pose estimation tasks [8] . It offers a strong trade-off between accuracy and computational cost, critical for detailed frame-by-frame diving analysis. As shown in Table II, under consistent bounding box conditions with RTMDet-nano, RTMPose-m provides robust performance across key evaluation metrics.

First, the diver's bounding box was detected using the RTMDet model, specifically the `rtmdet_m_640-8xb32_coco-person` configuration pre-trained on the COCO-person dataset. Within these detected bounding boxes, keypoints were extracted using RTMPose with the `rtmpose-m_8xb256-420e_body8-384x288` configuration. By suppressing background noise, the RVM-based input allowed RTMDet to more reliably localize the diver and reduced false positives. As a result, RTMPose received cleaner and more focused regions of interest, improving the precision of keypoint predictions, particularly around complex limb movements and rotations, as shown in figure 1.

TABLE II. PERFORMANCE COMPARISON OF RTMPOSE CONFIGURATIONS FOR BOUNDING BOX DETECTION CONFIGURATION - RTMDET_NANO_320-8XB32_COCO-PERSON ON INPUT VIDEO OF 961 FRAMES AND 38.44 SECONDS

Model Type	Input Res.	Dataset/Strategy	FPS	Total Time (s)
rtmpose-m	256x192	body8	5.98	160.77
rtmpose-l	256x192	body8	4.71	203.93
rtmpose-m	384x288	body8	5.07	189.54
rtmpose-l	384x288	body8	3.85	249.56
rtmpose-t	256x192	body8	6.93	138.73
rtmpose-s	256x192	body8	6.34	151.64
rtmpose-m	256x192	crowdpose	7.02	136.81
simcc_res50	384x288	simcc/coco	2.46	391.41
td-hm_hrnet	256x192	topdown/crowdpose	4.53	212.08
td-hm_res101	320x256	topdown/crowdpose	4.40	218.26

Our evaluation of the pose estimation results was primarily qualitative, focusing on the visual accuracy and consistency of the detected keypoints in various diving movements and poses. This synergy between accurate segmentation and efficient pose estimation allows for a detailed temporal analysis of the diver's articulated movements without excessive computational overhead. The choice of these specific RTMDET and RTMPose models enables us to achieve a balance between the need for precise keypoint localization in dynamic scenes and the

practical constraints of processing a large volume of video data.

F. KPI Calculation

We compute a set of KPIs essential for dive analysis: joint angles, total rotation, diver height, trajectory, and identification of dive phases. These metrics offer actionable insights for coaches and athletes, improving both training quality and performance assessment.

1) Joint Angles: The angle between three points A, B, and C in a 2D plane is calculated using the cosine rule, as expressed in Equation 1:

$$\theta_{ABC} = \arccos\left(\frac{(\mathbf{a}-\mathbf{b})\cdot(\mathbf{c}-\mathbf{b})}{\|\mathbf{a}-\mathbf{b}\|\|\mathbf{c}-\mathbf{b}\|}\right) \tag{1}$$

where $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are the position vectors of points A, B, and C respectively. To reduce the impact of noise in the pose estimation, the calculated joint angle time series were smoothed using a Kalman filter [16]. The Kalman filter is an optimal recursive estimator that predicts and updates the state (here, the joint angles) based on noisy measurements. The specific joint angles analyzed, along with the body parts used for their calculation, are listed below (referencing Equation 1):

- Torso Angle: Head, Left Shoulder, and Left Hip.
- Hip Angle: Left Shoulder, Left Hip, and Left Knee.
- Knee Angle: Left Hip, Left Knee, and Left Ankle.
- Arm Angle: Left Shoulder, Left Elbow, and Left Wrist.

2) Torso Orientation: The absolute orientation of the torso (ϕ_{Torso}) in the 2D plane was computed using the $\arctan 2$ function:

$$\phi_{\text{Torso}} = \arctan 2(y_{\text{Left Hip}} - y_{\text{Left Shoulder}}, x_{\text{Left Hip}} - x_{\text{Left Shoulder}}) \tag{2}$$

This angle, initially in radians, is converted to degrees for subsequent analysis:

$$\phi^{\circ}_{\text{Torso}} = \frac{180}{\pi}\phi_{\text{Torso}} \tag{3}$$

3) Total Rotation: The total rotation ($R(T)$) up to frame T is the cumulative sum of the absolute change in the torso orientation between consecutive frames, accounting for angle wrapping:

$$R(T) = \sum_{i=1}^{T} |\Delta\phi^{\circ}_{\text{Torso}}(t_i)| \tag{4}$$

where $\Delta\phi^{\circ}_{\text{Torso}}(t_i)$ is the change in torso orientation in degrees between frame $i-1$ and i, with adjustments made for angles crossing the $\pm 180^{\circ}$ boundary. This total rotation data is subsequently smoothed by using a Gaussian filter to reduce high-frequency noise.

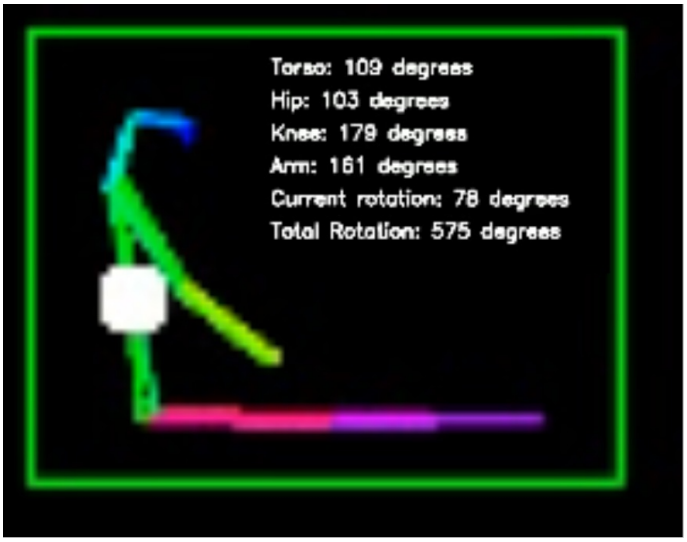

Fig. 2. Visual representation of angle calculations following equations 1, torso orientation (ϕ_{Torso}) and cumulative total rotation ($R(T)$).

4) Diver Height (Center of Mass Y-coordinate): The vertical position of the diver is tracked using the y coordinate of the Center of Mass (y_{COM}), calculated as the midpoint between the Left Hip and Left Shoulder:

$$y_{\text{COM}} = \frac{y_{\text{Left Hip}} + y_{\text{Left Shoulder}}}{2} \tag{5}$$

The maximum height reached by the diver (H_{max}) during the dive is then:

$$H_{\text{max}} = \max_{t}\{y_{\text{COM}}(t)\} \tag{6}$$

5) Diver Height (Real-World Y-coordinate): To obtain the real-world height of the diver, we first establish a scaling factor based on known dimensions in the initial frame where the diver is on the board. The scaling factor is calculated as:

$$s = \frac{H_{\text{board}} + H_{\text{diver, initial}}}{P_{\text{diver, board}} - P_{\text{water}}} \tag{7}$$

where:

- s is the scaling factor (meters per pixel).
- H_{board} is the height of the diving board above the water in meters.
- $H_{\text{diver, initial}}$ is the approximate height of the diver while standing on the board in meters.
- $P_{\text{diver, board}}$ is the pixel y-coordinate of a reference point on the diver (e.g., the feet) while on the board.
- P_{water} is the pixel y-coordinate of the water level.

The real-world y-coordinate of the Center of Mass (Y_{COM}) in meters is then calculated by converting the pixel-based y-coordinate (y_{COM}) using the scaling factor and adjusting for the board height:

$$Y_{\text{COM}}(t) = y_{\text{COM}}(t) \cdot s - H_{\text{board}} \tag{8}$$

The maximum height reached by the diver ($H_{\text{max, meters}}$) relative to the water level during the dive is:

$$H_{\text{max, meters}} = \max_t \{Y_{\text{COM}}(t)\} \tag{9}$$

Similar to the total rotation data, the time series data for the diver's real-world y-coordinate (and consequently the maximum height) is smoothed using a Gaussian filter to reduce temporal noise.

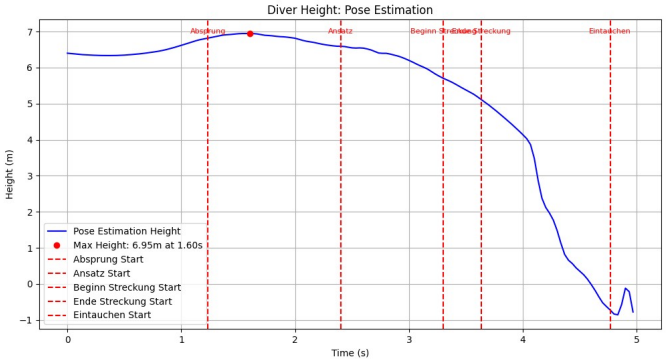

Fig. 3. Visual representation of (Y_{COM}) over entire dive period. The maximum height ($H_{\text{max, meters}}$) is indicated in big red dot along with different dive stages indicated in dashed lines.

III. Results

This section presents the quantitative results obtained from applying the proposed pose estimation pipeline to a 205B dive: Forward 2 1/2 Somersault Pike.

A. Joint Angles

Figure 4 shows the temporal evolution of the filtered joint angles (Torso, Hip, Knee, and Arm) throughout the dive. Dive stage events are indicated on the plots. We observe distinct patterns of flexion and extension in each joint corresponding to the different phases of the dive. For instance, the knee angle shows a sharp increase during the takeoff phase followed by a decrease during the somersault.

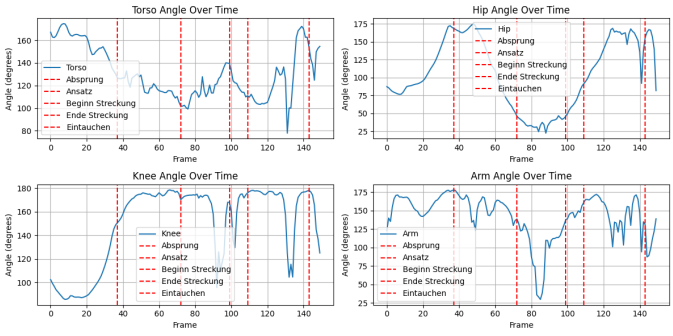

Fig. 4. Filtered Joint Angles Over Time for a Forward 2 1/2 Somersault.

B. Torso Orientation and Total Rotation

The absolute torso orientation throughout the dive is depicted in Figure 5. The total rotation calculated for this dive was 890 degrees, the intersection of Eintauchen and total rotation at approximately 890 degrees.

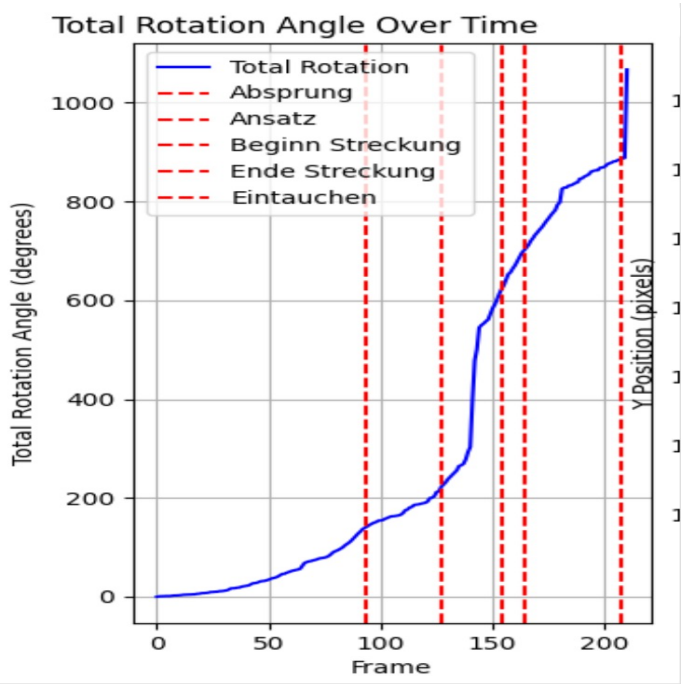

Fig. 5. Torso Orientation Over Time for a Forward 2 1/2 Somersault Dive.

C. Diver Height Profile and Dive Phases

Figure 3 shows the real-world height of the diver's center of mass over the duration of the dive. The maximum height reached was 6.95 meters at 1.60s (indicated in red dot).

IV. Discussion

The primary goal of this study was to develop and evaluate a pose estimation pipeline for detailed analysis of sports diving. The results obtained from analyzing a forward 2 1/2 somersault pike (205B) dive demonstrate the pipeline's capability to extract key performance indicators such as joint angles, torso orientation, total rotation, and diver height.

A. Comparison with Manual Calculations

The total rotation calculated by our pipeline for the example dive was 890 degrees, compared to a manual calculation of 882 degrees. This difference of 8 degrees could be attributed to discrepancies during complex dive stages and occlusion during dive end stage. Similarly, the maximum height achieved by the diver's center of mass was measured as 6.95 meters by the pipeline and 6.89 meters manually (5 meters of board height, 1.7 meters of diver height and 0.19 manual maximum height calculation). The reasons for these variations warrant further investigation, potentially focusing on occlusion caused by lighting, water splashing, complex body rotations.

B. Biomechanical Interpretation of the 205B Dive

The temporal profiles of the joint angles (Figure 4) show patterns consistent with the biomechanics of a forward 2 1/2 somersault pike dive. The flexion at the hips and knees during the pike position and the extension during takeoff and entry are clearly visible. The total rotation of 890 degrees aligns with the expected 900 degrees for a 2 1/2 somersault, with the observed trajectory in Figure 5 indicating a consistent rotational velocity throughout the aerial phase. The maximum height of 6.95 meters relative to the water surface, as measured by our pipeline, shows good agreement with a manual calculation that estimates a maximum height of 0.19 meters above the diver's initial standing height of 1.7 meters on a 5-meter board, resulting in a total height of 6.89 meters. Achieving adequate height is crucial for a complex dive like the 205B, providing the necessary time for the 2 1/2 somersaults and the execution of the pike position before entry. The slight overestimation of 0.06 meters by our pipeline could be attributed to minor variations in the defined center of mass or frame-level inaccuracies in pose estimation. Overall, the maximum height measurement indicates an effective takeoff for this dive.

C. Pipeline Performance

The performance of the dive analysis pipeline was evaluated on a video sequence of a forward 2 1/2 somersault pike dive, spanning 6 seconds (from 12.0 to 18.0 seconds) and comprising 179 frames at an original frame rate of 30 FPS. The processing was conducted on a local machine equipped with an NVIDIA RTX 3060 GPU and an Intel Core i7-11800H CPU.

Table III summarizes the processing time and average frames per second (FPS) for each key stage of the pipeline for this 179-frame segment.

TABLE III. PIPELINE PERFORMANCE BREAKDOWN (179 FRAMES)

Stage	Total Time (s)	Average FPS
Preprocessing	11.69 (for 1064 frames)	91.02
Background Removal (RVM)	25.63	6.98
Pose Estimation (RTMPose)	93.22	1.92
Keypoint Visualization	3.27	54.78

As the table illustrates, the pose estimation stage is the most computationally intensive, representing the current bottleneck in the pipeline with an average speed of 1.92 FPS. The background removal using RVM achieves a significantly higher frame rate of 6.98 FPS. The keypoint visualization is the fastest stage, running at near real-time speeds. The effective end-to-end processing speed for the core analysis steps (segmentation and pose estimation) for this video segment is approximately 1.51 FPS. While this allows for detailed offline analysis and KPI extraction, future work could explore optimizations in the pose estimation parameters or hardware acceleration to improve the real-time capabilities of the system.

D. Limitations

This study presents a detailed analysis of a single dive type. Further validation across a wider range of dives and athletes is necessary to fully assess the generalizability of the pipeline. The qualitative evaluation of the pose estimation accuracy provides initial confidence, but a quantitative evaluation using ground truth data would be beneficial in future work. Additionally, the current reliance on a fixed scaling factor could be improved by incorporating dynamic calibration techniques.

V. SUMMARY AND OUTLOOK

This work demonstrated that AI-based pose estimation can enable detailed, frame-level analysis of diving performance. The proposed pipeline successfully tracked body keypoints and extracted key performance indicators such as joint angles, rotation, and maximum height, that support automated evaluation with high efficiency and accuracy.

Currently, dive stage detection is limited to only somersault dive types. Future work can extend this to other dive types by training a custom Temporal Convolutional Neural Network (TCNN) or other deep learning models tailored for dynamic stage classification. Broader goals include improving robustness, enabling real-time feedback, and supporting applications such as automated judging and injury prevention.

The key contribution of this work is to show that accurate, high-speed pose tracking in a complex, dynamic sport such as diving is not only feasible but also scalable.

ACKNOWLEDGMENT

The authors would like to thank Alexander Weiß, Researcher and PhD Candidate at Lehrstuhl für Maschinelles Lernen und Datenanalytik for mentoring the project.

REFERENCES

[1] Fukushima, T., Blauberger, P., Guedes Russomanno, T., Stöveken, J.-H., and Eskofier, B.: *The potential of human pose estimation for motion capture in sports: a validation study*. Sports Eng 27, 19 (2024), doi: 10.1007/s12283-024-00460-w.

[2] Badiola-Bengoa, A., and Mendez-Zorrilla, A.: *A Systematic Review of the Application of Camera-Based Human Pose Estimation in the Field of Sport and Physical Exercise.* Sensors (Basel, Switzerland), 21(18), 5996 (2021), doi: 10.3390/s21185996.

[3] Grannan, C.: *How Is Diving Scored? Encyclopedia Britannica.* https://www.britannica.com/story/how-is-diving-scored. Last visited: 28.03.2025 (2025).

[4] Jakab, S., Davis, P., and Whyte, I.: *exploratory investigation of traditional scoring in diving and relationships to the development of Artificial Intelligence opportunities.* Scientific Journal of Sport and Performance, 2(1), 300-313 (2023), doi: 10.55860/QELM3130.

[5] Murthy, P., Taetz, B., Lekhra, A., and Stricker, D.: *DiveNet: Dive Action Localization and Physical Pose Parameter Extraction for High Performance Training*. IEEE Access 11, 37749–37767 (2023), doi: 10.1109/ACCESS.2023.3265595.

[6] Park, J., and Yoon, S.: *Kinematic analysis of back somersault pike according to skill level in platform diving.* Korean Journal of Sport Biomechanics, 27 (2017), 157–164. doi:10.5103/KJSB.2017.27.3.157.

[7] Lin, S., Yang, L., Saleemi, I., and Sengupta, S.: *Robust High-Resolution Video Matting with Temporal Guidance.* https://arxiv.org/abs/2108.11515

[8] Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., and Chen, K.: *RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose.* https://arxiv.org/abs/2303.07399

[9] Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., and Chen, K.: *RTMDet: An Empirical Study of Designing Real-Time Object Detectors.* https://arxiv.org/abs/2212.07784

[10] Kirillov, A. et al.: *Segment Anything.* https://arxiv.org/abs/2304.02643. Last visited: 28.03.2025 (2025)

[11] Google AI, MediaPipe Selfie Segmentation. https://ai.google.dev/edge/mediapipe/solutions/vision. Last visited: 28.03.2025 (2025)

[12] Glenn Jocher, YOLOv5. https://github.com/ultralytics/yolov5. Last visited: 28.03.2025 (2020)

[13] Google AI, MediaPipe Repository. https://github.com/google-ai-edge/mediapipe/tree/master. Last visited: 28.03.2025 (2025)

[14] Google AI, MoveNet - SinglePose Thunder. https://www.kaggle.com/models/google/movenet/tensorFlow2/singlepose-thunder. Last visited: 28.03.2025 (2025)

[15] Jo, B and Kim, S.: *Comparative Analysis of OpenPose, PoseNet, and MoveNet Models for Pose Estimation in Mobile Devices.* Traitement du Signal, 39 (2022), 119-124. doi:10.18280/ts.390111.

[16] Kalman, R. E.: *A new approach to linear filtering and prediction problems.* Journal of basic Engineering, 82(1) (1960), 35-45.