Measurement: Digitalization 6 (2026) 100029

Contents lists available at ScienceDirect

Measurement: Digitalization

journal homepage: www.elsevier.com/locate/meadig

Full Length Article

Problem-related performance metrics of deep learning models: Application to swimmer pose estimation in underwater environments

Alessia Caputo *, Alberto Scocco, Paolo Castellini

Department of Industrial Engineering and Mathematical Sciences, Università Politecnica delle Marche, Via Brecce Bianche 12, 60131, Ancona, Italy

ARTICLE INFO

Keywords:
Underwater swimmer pose estimation
Problem-oriented performance metrics
Scale-consistent anatomical tolerances
Continuous localization accuracy
Multi-scale keypoint evaluation
Sports biomechanics.

ABSTRACT

Accurate 2D swimmer pose estimation in underwater environments remains a challenging task due to optical distortions, dynamic occlusions, and the highly multi-scale nature of anatomical landmarks. Conventional evaluation metrics adopted from generic computer vision benchmarks are often inadequate to characterise the functional reliability required in sports biomechanics, particularly when small and fast-moving joints are involved.

This work proposes a problem-orientated evaluation framework for underwater swimmer pose estimation, applied to seven deep learning model configurations differing in training data composition, pre-processing strategies, and parameter optimisation. Beyond the indicators based on standard confusion-matrix, a dual assessment strategy is introduced, combining strict anatomical tolerance thresholds with a continuous tolerance-Normalised Localisation Accuracy (NLA). Keypoint-specific tolerances are derived from the spatial extent of each anatomical region, allowing scale-consistent evaluation throughout the kinematic chain.

Experimental results show a pronounced performance gradient from core body segments to distal extremities, highlighting the limitations of binary metrics for small joints. Models trained on heterogeneous raw datasets achieve the best overall performance (Global Performance Index = 78.69), demonstrating superior robustness and generalisation.

Comparative analysis reveals that binary tolerance-based metrics are overly punitive for distal landmarks and tend to obscure the true localisation capability of the models. The proposed continuous NLA provides a more informative representation of spatial uncertainty and measurement quality. These findings emphasise the importance of problem-related, scale-aware evaluation metrics and confirm data diversity as a more effective driver of robustness than aggressive pre-processing in underwater swimmer pose estimation.

1. Introduction

Human Pose Estimation (HPE) from monocular RGB images has become a central challenge in computer vision, enabled by deep neural networks and supported by large annotated datasets. Modern 2D human pose estimators that achieve a “good enough” accuracy in specific tasks are routinely applied in human-computer interaction, surveillance, healthcare, and sports biomechanics [1,2]. In recent years, convolutional architectures and multi-stage refinement models have substantially improved the robustness of 2D pose estimation. Representative examples include the classic OpenPose and its Part Affinity Fields formulation for multi-person pose in challenging scenes [3], and detection-based pipelines derived from general-purpose object detectors such as YOLO [4]. More recent are solutions such as the ViTPose family, which propose an approach based on the transformer architecture applied to computer vision [5], adding flexibility and scalability to pose detection solutions. Markerless HPE offers a non-invasive way to capture kinematic data, enabling the assessment of peak performance and technique without the limitations of wearable biomechanical sensors.

Several challenge factors can reduce the reliability of markerless tracking and HPE, particularly in aquatic environments. Visual data can be hidden or distorted by water turbulence, bubbles, and complex light refraction at the air-water interface, leading to high uncertainty in keypoint estimation [6–8].

In complex movements, human keypoints are often hidden by the athlete’s own body or by swimming pool equipment [9], and in addition the dynamic lighting and “cluttered” backgrounds common in high-performance settings can confuse Deep Learning (DL) models during the crucial pixel-to-keypoint mapping process [10].

Several works have addressed aquatic sports to improve detection. Some explored architectures based on stroke-type conditioning [11], while others investigated pose estimation in surveillance-like scenarios at pool borders [12]. Dedicated frameworks have also been introduced, using fully convolutional neural networks adapted to the visual

* Corresponding author.
E-mail addresses: a.caputo@pm.univpm.it (A. Caputo), a.scocco@pm.univpm.it (A. Scocco), p.castellini@univpm.it (P. Castellini).

https://doi.org/10.1016/j.meadig.2026.100029
Received 27 December 2025; Received in revised form 6 March 2026; Accepted 11 March 2026
Available online 15 March 2026
3050-6441/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

characteristics of submerged athletes [13].

These methods usually form the backbone of many downstream systems, including sports analytics frameworks that extract kinematic information from picture or video recordings. Their evaluation is still largely driven by generic computer vision benchmarks, such as the MPII Human Pose and COCO keypoints datasets, where acquisition conditions, visibility patterns, and error budgets differ significantly from those encountered in underwater sports applications [14,15].

Standard classification metrics derived from the confusion matrix, such as Accuracy, Precision, Recall, or the F1/Dice Score, are effective for characterising the model's ability to detect presences or handle occlusions. Detections are classified into the four disjoint categories that characterise the confusion matrix:

- TP (True Positives): Correctly predicted positive cases;
- FP (False Positives): Incorrectly predicted as positive (Type I error);
- TN (True Negatives): Correctly predicted negative cases;
- FN (False Negatives): Incorrectly predicted as negative (Type II error).

Then, counting each single case and result given by a model, it is possible to evaluate these quality measures:

• $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	This measures overall correctness and is evaluated as the proportion of correct predictions out of all predictions. However, accuracy can be misleading with imbalanced datasets.
• $Precision = \frac{TP}{TP+FP}$	It measures how reliable your positive predictions are.
• $Recall = \frac{TP}{TP+FN}$	It measures completeness or sensitivity and high recall means that the system is catching most positive cases.
• $Dice\ Score = \frac{2 \cdot TP}{2 \cdot TP+FP+FN}$	This is the harmonic mean of precision and recall, combining and balancing both into one metric.

These metrics help characterise the model used, for example, in the following cases:

- Detecting whether a keypoint is visible or occluded (binary);
- Classifying the action or pose category;
- Detecting whether a person is present before estimating their pose.

This can be done in a generic context, but the specific case of underwater swimmer analysis, acquisition conditions, visibility patterns, and error budgets differ significantly from terrestrial scenarios. Underwater visual sensing introduces specific degradations: light propagation is governed by wavelength-dependent absorption and scattering which, together with non-uniform lighting, reduce contrast and signal-to-noise ratio [16]. Although image enhancement methods and generative approaches have been proposed to restore useful content [17,18], pose estimation in the underwater context of a swimmer's pose from a submerged side view raises unique challenges:

- **Side-view Occlusions:** Due to the lateral camera position, keypoints on the swimmer's contralateral side are frequently occluded by the body itself;
- **Air-Water Interface:** some keypoints exit the water during the stroke recovery phase, becoming invisible or heavily distorted by refraction (e.g. finger and foot tips);
- **Environmental Noise:** The presence of air bubbles and turbulence can obstruct the view of keypoints or create false positives;
- **Dynamic Constraints:** The high speed of distal joints (hands, feet) combined with their small apparent size makes precise localisation difficult compared to the torso;
- **Non-canonical Views:** In the swimming pose analysis we are considering, the body is always in a lateral horizontal view, which is quite unusual for the general purpose training made in pose detection models.

This complex scenario suggests that swimming pose estimation is a hybrid problem, lying between pure classification (determining if a keypoint is visible or not) and regression (identifying the coordinates of visible keypoints).

Considering the problem in its regression aspect, when it is necessary to predict continuous coordinates (x, y, and sometimes z) for keypoints like joints, generally different metrics are used:

- PCK (Percentage of Correct Keypoints): A keypoint is considered "correct" if it falls within a normalised distance threshold from the ground truth;
- OKS (Object Keypoint Similarity): It accounts for keypoint visibility and applies per-keypoint standard deviations.
- MPJPE (Mean Per Joint Position Error): The average Euclidean distance between predicted and ground truth keypoints, measured in millimetres or pixels for each frame, and then averaged.
- PA-MPJPE (Procrustes-aligned MPJPE): MPJPE after aligning the predicted pose to ground truth using rotation, translation, and sometimes scaling;
- PCKh (Normalised PCK): measures the percentage of predicted keypoints that fall within a certain distance (threshold) of the ground-truth location [19].

The distance is typically normalised by a specific body measurement to account for variations in person size and viewpoint in the images. Such kinds of transformation usually consider the person's bounding box size (e.g. the diagonal of the box [19]), or by head size (usually the distance between eyes). There is no difference in normalisation of different size keypoints. Furthermore, when thresholds are used, they are common and of a given size, so the detection of a key point is considered "correct" if the prediction falls within the area of interest (e.g. 20% of the size of the bounding box relative to the reference data).

Going further, the Scale Consistency Score (SCS) [20,21] is an evaluation metric specifically designed in this research for multi-scale scenarios, related to the core objective of the problem, but it is not a single universal standardisation metric. On the other hand, Average Precision (AP) evaluates model performance from the perspective of an anatomical structure of the body. It reflects the fine granularity the detection difficulty of different joints and the capability of the model, but does not take into account the "relevance" of a given error when compared to the characteristic size of the keypoint.

These kind of regression metrics are effective to compare general-purpose models, and provide compact accuracy indicators that aggregate performance over all joints and images, but in the swimming sports context it is important to quantify the spatial precision of the prediction and not mask the specific failure modes. Moreover, it is relevant for human pose analysis in sports applications to distinguish between a slight inaccuracy and a gross error, considering the different functional importance of each anatomical segment.

Despite this progress, the definition of evaluation metrics that reflect the needs of sports biomechanics remains an open issue. In particular, the assessment of swim-technique depends on how accurately each anatomical landmark is localised relative to its specific spatial extent, whether predictions remain stable over time, and how the model behaves in the presence of legitimate occlusions.

The proposed study addresses these issues through a methodological evaluation of the 2D underwater swimmer pose estimation. Seven model configurations are analysed and compared, differing in the dataset composition for the training phase, the background subtraction strategy, and parameter optimisation. To overcome the limitations of standard metrics, a custom evaluation pipeline is introduced. This pipeline refers to ground-truth annotations of 17 anatomical keypoints, keypoint-specific

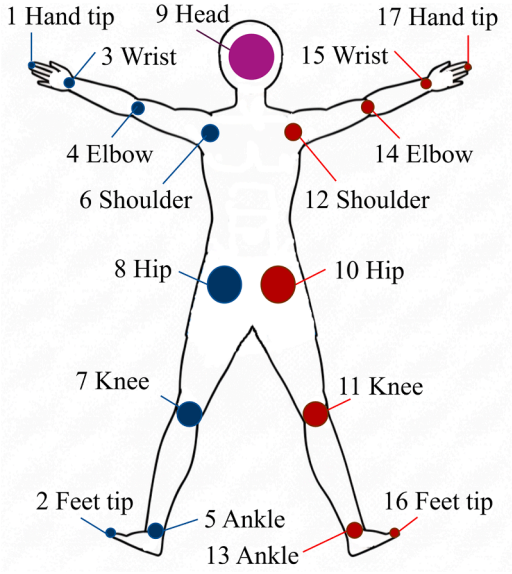

Fig. 1. Schematic of the 17 anatomical keypoints annotated for the swimmer pose estimation task.

tolerance thresholds derived from their spatial extent, and explicit visibility masks. This research aims to provide quantitative support to select a training strategy and allows a direct comparison between traditional and metric sensitive evaluation approaches to the functional tolerance of swimmer biomechanics.

2. Materials and methods

2.1. Dataset and anatomical annotation protocol

The experimental dataset consists of underwater video sequences of competitive athletes acquired in a swimming pool environment. Lateral cameras were positioned to capture the entire stroke cycle. The footage includes freestyle, breaststroke, and butterfly trials, recorded under comparable illumination and acquisition settings. Individual video sequences corresponding to a single swimmer and stroke condition were processed as separate recordings. The dataset used for model development comprises recordings of 20 unique competitive swimmers, each contributing multiple underwater video sequences across the different stroke types.

For each recording, a subset of frames was manually annotated with 2D image coordinates of $N = 17$ anatomical landmarks. The annotation protocol defines a central head landmark and symmetric pairs on the upper and lower extremities: shoulder, elbow, wrist, hand, hip, knee, ankle, and feet, both for the left and right side (Fig. 1).

All underwater video sequences used in this work were collected specifically for this study. Ground-truth annotations were produced manually using a custom point-and-click GUI tool to support consistent and accurate labelling. Landmarks that are outside the field of view, heavily occluded, or not reliably visible were explicitly flagged as missing.

Fig. 2 shows a representative frame extracted from the underwater acquisition dataset, with 13 of the 17 annotated anatomical keypoints used in this study.

Each annotation file contains one row per landmark and frame, which contains horizontal and vertical pixel coordinates, while landmarks that are outside the field of view, heavily occluded, or not reliably visible are explicitly flagged as *missing*. These annotations are excluded from localisation error computations, but are essential for evaluating the visibility classification performance (True/False Negatives).

To assess intra-annotator consistency, the same operator was asked to repeatedly annotate the pelvic keypoint on a representative underwater frame using the custom GUI. Over 20 repeated clicks, the radial distance of the annotations from their mean position showed an average value of 2.8 px with a standard deviation of 1.5 px, indicating that the variability of manual labelling is limited to a few pixels and markedly smaller than the keypoint-specific tolerance radii used in the subsequent analysis.

2.2. Pose estimation model configurations

To investigate the efficacy of the proposed metric, seven models were implemented using a hybrid ensemble that combines two state-of-the-art convolutional HPE solutions. Ultralytics YOLOv8 was used for the detection of the main body keypoints, while Google Mediapipe v0.10.7 was used to complement the detection of hand and foot tips. To test the metric under different conditions, the models were trained using varying combinations of athletes, swimming styles, and sizes of the training and test datasets, as well as training hyperparameters such as the number of epochs and the learning-rate schedule. These differences intentionally produced a variety of performance levels, allowing for a meaningful assessment of the behaviour of the proposed measure. The configurations are defined as follows:

- **Model 1:** Trained on a heterogeneous dataset of underwater videos featuring three strokes (freestyle, breaststroke, butterfly). The total training set consists of approximately 8.800 frames.
- **Model 2:** Trained on a restricted subset of underwater videos containing only freestyle sequences. The dataset consists of 8.000 frames.
- **Model 3:** Trained on freestyle-only videos. The total number of frames used is 14.000 frames.
- **Model 4:** Uses the same training data as Model 3, but with optimised hyper-parameters to improve convergence.
- **Model 5:** Trained on the same training frames used for Model 4, but processed with a medium-level background subtraction algorithm to remove static noise (e.g. tiles, lane ropes).
- **Model 6:** Uses the same pre-processed videos as Model 5, combined with the optimised parameter settings of Model 4.
- **Model 7:** Builds upon Model 6 by applying regularisation techniques (e.g. early stopping, data augmentation) to mitigate overfitting observed in the most data-rich configurations.

Pose estimation models were applied in offline batch mode, processing individual frames as independent static images rather than continuous video streams. Real-time performance was not required in this study and therefore inference speed (FPS) was not systematically characterised. All HPE models and the MATLAB scripts used to compute the performance metrics were executed on the same workstation with the following configuration: CPU: Intel Core i7-11700 (64 GB RAM); GPU: 2× Nvidia RTX 3090 (24 GB VRAM each); OS: Linux Ubuntu 22.04 LTS (64-bit) with Nvidia drivers 550.163.01.

All models produce a prediction file with the same structure of ground-truth annotations, containing the 2D coordinates of all landmarks or a missing entry NaN NaN when no prediction is generated.

The composition of training and testing datasets was managed using K-fold cross-validation with an 80/20 split. For each model, the frames used for training and testing were drawn from the subset of recordings specifically assigned to that configuration, so the corresponding training and test sets contained images taken from the same athletes.

For the comparative analysis reported in Results and Discussion session, all performance metrics are computed on an additional test sequence acquired from a swimmer who is not included in the training datasets of the analysed models. This evaluation strategy reduces the risk of overfitting to individual body shapes while keeping the training protocol focused on the effect of underwater image quality, distortions, and pre-processing choices.

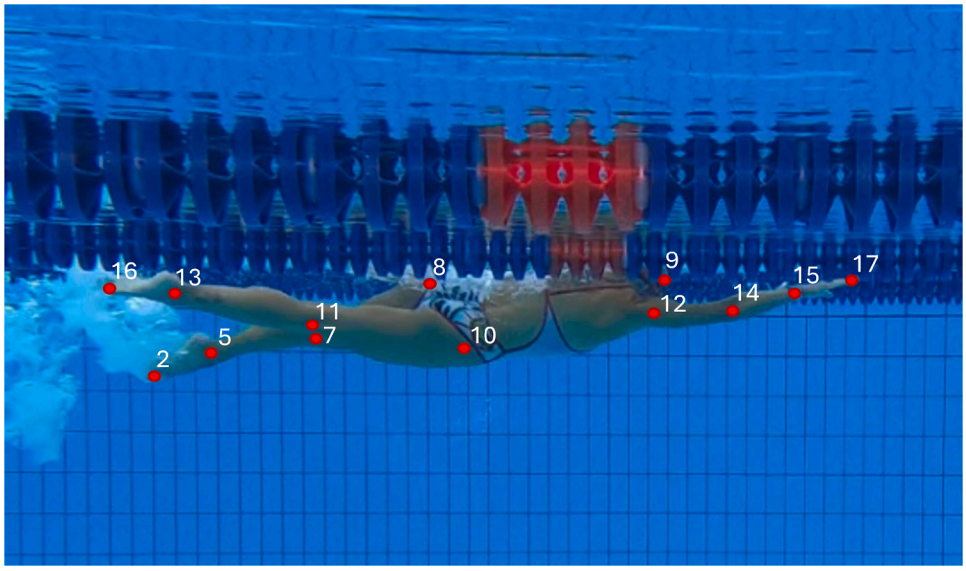

Fig. 2. Representative underwater frame with 13 of the 17 anatomical keypoints.

Since each anatomical landmark covers different areas of the image and has a specific functional role, applying a uniform spatial tolerance is inappropriate. To address this, a representative frame was selected to calibrate a characteristic circular Region of Interest (RoI) for each keypoint. Assuming that the circular area A_i (in px^2) encloses the anatomical feature, a specific tolerance parameter can be defined for the keypoint i as the radius of the circular RoI and can be derived as:

$$r_i = \sqrt{\frac{A_i}{\pi}} \tag{1}$$

This formulation ($r_i \propto \sqrt{A_i}$) establishes a relation between the admissible localisation error and the apparent size of the anatomical region (Fig. 3): larger segments (e.g. hips) allow for larger deviations, while smaller and localised elements (such as wrists or hand tips) require tighter tolerance.

2.3. Frame-wise categorisation and continuous localisation measure

For every model, ground-truth (GT) annotations and predicted coordinates are compared frame-by-frame: where a landmark is not visible, it is treated as negative for that keypoint, whereas a frame with a valid annotation is considered positive.

Each keypoint and frame are assigned to one of four elementary categories:

- True Negative (TN): keypoint is not annotated (not visible), and no prediction is produced;
- False Positive (FP): keypoint is not annotated, but the model predicts a coordinate;
- False Negative (FN): keypoint is annotated (visible), but no prediction is produced;
- True Positive (TP): keypoint is annotated, and a prediction is available.

To provide a strict assessment of detection validity, consistent with anatomical constraints, the percentage of True Positives in Tolerance was computed. In this binary classification, a predicted keypoint is considered valid (1) only if its Euclidean distance $d_{i,f}$ from the ground truth lies strictly within the radius r_i, and otherwise invalid (0) .

$$Valid_{i,f} = \begin{cases} 1 & \text{if } d_{i,f} \le r_i \\ 0 & \text{if } d_{i,f} > r_i \end{cases} \tag{2}$$

This metric corresponds to a PCK evaluation with a strict anatomical approach. However, explicitly excluding predictions located just outside the threshold (e.g. $d = 1.01 \cdot r_i$) fails to differentiate between a slightly inaccurate detection and a gross localisation error. To address this limitation, a continuous scoring metric is subsequently introduced. For True Positive cases, the localisation quality is quantified using the Euclidean distance $d_{i,f}$ between the predicted $\mathbf{p}_{i,f}$ and ground-truth $\mathbf{g}_{i,f}$ coordinates. To derive a bounded quality metric suitable for aggregation, we define the NLA based on a normalised and saturated error. First, the error is normalised by the keypoint-specific radius r_i and saturated at a threshold 1:

$$e^{\text{sat}}_{i,f} = \min\left(\frac{d_{i,f}}{3r_i}, 1\right) \tag{3}$$

The saturation threshold at $3r_i$ is introduced to define a biomechanically meaningful upper bound. Since r_i approximates the characteristic spatial extent of the anatomical region, an error exceeding three times this radius indicates that the predicted keypoint is no longer spatially compatible with the underlying anatomical structure. In such cases, the localisation becomes functionally unusable for joint-angle computation and downstream biomechanical parameter extraction. The factor 3 was selected as a conservative limit based on expert consultation, who identified this deviation as beyond acceptable tolerance for technique assessment. Errors larger than this threshold are therefore saturated in the score in order to prevent a disproportionate influence of gross outliers on the aggregate metric (Fig. 4). The continuous NLA $S_{i,f}$ (inside the range $[0, 100]$) is then computed as a linear mapping of the saturated error:

$$S_{i,f} = 100 \cdot \left(1 - e^{\text{sat}}_{i,f}\right) \tag{4}$$

According to this definition, a perfect location ($d = 0$) yields a score of 100, while a localisation at the tolerance limit ($d = r_i$) yields 66.7, and any error exceeding $3r_i$ yields a score of 0.

2.4. Performance metrics and aggregate indices

To evaluate the reliability of the model from a biomechanical perspective, raw detection results are converted to standardised performance metrics. All spatial metrics are normalised to keypoint-specific tolerances: localisation errors are scaled by the radius $r_i \propto \sqrt{A_i}$, ensuring that a spatial deviation on a small joint (e.g. wrist) is penalised more heavily than an equivalent pixel error on a large segment (e.g. hip). Four

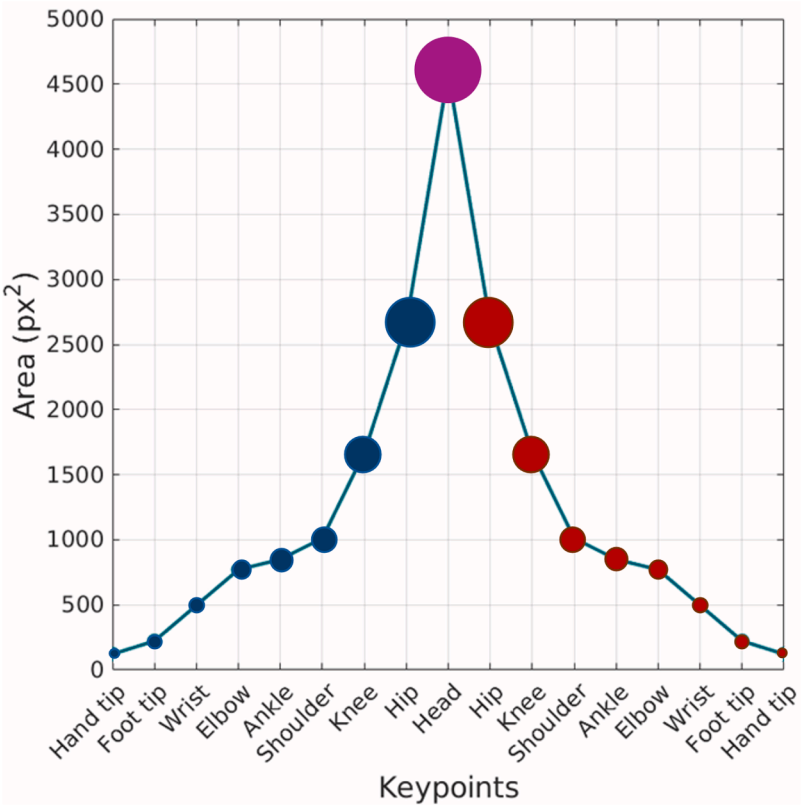

Fig. 3. Keypoints vs detection area sizes (px^2).

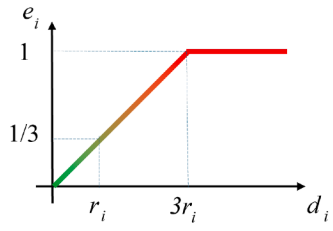

Fig. 4. Linear progression of the error till the ($3r_i$) threshold.

complementary indices (scaled $0-100$, where 100 represents optimal performance) are defined to characterise behaviour at specific anatomical keypoints and to compute the Global Performance Index (GPI) visualised in radar charts:

- **Normalized Localisation Accuracy** (S_{acc}): Represents the positional accuracy of correctly detected keypoints. It corresponds to the mean continuous score computed over all True Positive instances. A value of 100 implies perfect overlap with the ground truth.
- **Detection Sensitivity** (S_{sens}): Measures the ability to correctly identify a landmark when it is visible. It is defined as the complement of the FN:

$$S_{sens} = 100 - FN\%_{|GT_{pos}} \tag{5}$$

  High sensitivity corresponds to a low False Negative rate ($FN\%$).
- **Rejection Specificity** (S_{spec}): Quantifies the reliability in handling occlusions and out-of-view keypoints. It is defined as the complement of the FP:

$$S_{spec} = 100 - FP\%_{|GT_{neg}} \tag{6}$$

  High specificity corresponds to a low False Positive rate ($FP\%$), indicating robustness against hallucinations.

For completeness, three standard pose-estimation metrics commonly used in the literature were also computed on the independent test sequence adopted in this study: MPJPE, PCK and OKS.

Formally, let $d_{i,f}$ denote the Euclidean distance between predicted and ground-truth coordinates for keypoint i in frame f, and let $\mathcal{V}$ be the set of all keypoint–frame pairs for which both ground truth and prediction are available:

$$\mathcal{V} = \{(i, f) : \mathbf{g}_{i,f} \text{ and } \mathbf{p}_{i,f} \text{ are both defined}\}.$$

The Mean Per Joint Position Error (MPJPE) is defined as:

$$\text{MPJPE} = \frac{1}{|\mathcal{V}|} \sum_{(i,f)\in\mathcal{V}} d_{i,f}. \tag{7}$$

For PCK, let D_f denote the diagonal of the ground-truth bounding box enclosing all visible keypoints in frame f, and let the correctness threshold be $\tau_f = 0.2\, D_f$. The PCK score is then given by:

$$\text{PCK} = 100 \cdot \frac{\sum_{(i,f)\in\mathcal{V}} \mathbb{I}\left(d_{i,f} \le \tau_f\right)}{\sum_{(i,f)\in\mathcal{V}} 1}, \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

For OKS, let A_f be the area of the ground-truth bounding box in frame f, $s_f = \sqrt{A_f}$ a scale factor, and κ_i a fixed tolerance factor for keypoint i. The frame-wise OKS is defined as:

$$\text{OKS}_f = \frac{1}{|\mathcal{V}_f|} \sum_{i\in\mathcal{V}_f} \exp\left(-\frac{d_{i,f}^2}{2\,(s_f\kappa_i)^2}\right), \tag{9}$$

where $\mathcal{V}_f = \{i : (i, f) \in \mathcal{V}\}$, and the overall OKS is obtained by averaging over all frames where at least one keypoint is visible:

$$\text{OKS} = \frac{1}{|\mathcal{F}_{\text{OKS}}|} \sum_{f\in\mathcal{F}_{\text{OKS}}} \text{OKS}_f. \tag{10}$$

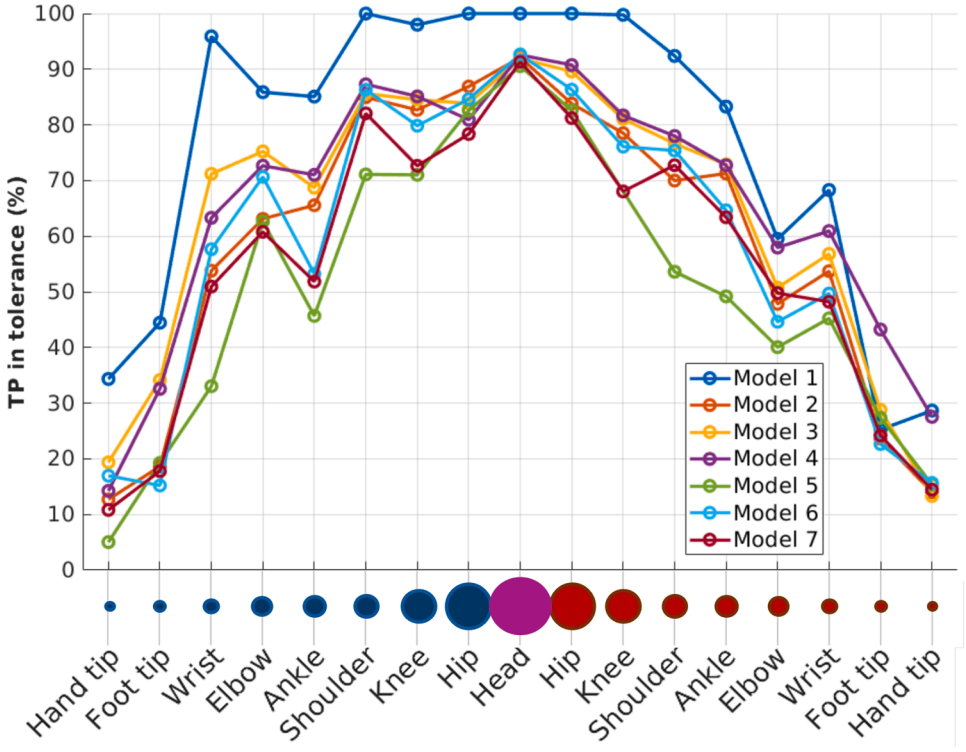

Fig. 5. Percentage of true positives within anatomical tolerance ($TP_{in}\%$) across all keypoints.

These metrics are then used in the Results section to provide a direct comparison between the proposed NLA-based indices and commonly adopted evaluation criteria.

3. Results

The proposed metrics were applied to compare the seven model configurations with the common test set. Results are organised to highlight, first, the per-keypoint localisation capability and, second, the distribution of detection outcomes conditioned on landmark visibility.

3.1. Anatomical tolerance analysis

Fig. 5 illustrates the percentage of True Positives within the anatomical tolerance ($TP_{in}\%$) for all models in the 17 keypoints. The graph exhibits a characteristic "bell-shaped" profile common to all configurations: central segments (Head, Hips) achieve high detection rates (near 100% for the best models), while performance degrades significantly towards the distal extremities.

Model 1 (dark blue line) consistently outperforms other configurations, particularly on the left-side extremities (hand tip, foot tip, wrist), where it maintains a detection rate between 35% and 95%, significantly higher than the background-subtracted models (Models 5, 6, 7), which drop below 20% for hand tips. This suggests that while strict anatomical thresholds are challenging for small, fast-moving joints, the model trained on the most heterogeneous dataset retains superior spatial precision.

3.2. Global performance assessment

To provide a holistic view of the trade-off between localisation precision and detection reliability, the four aggregate indices—NLA, Detection Sensitivity, Rejection Specificity and True Negative Rate—are visualised in the radar chart in Fig. 6.

The shape of the polygons reveals significant behavioural differences. Model 1, trained on the most heterogeneous dataset, maintains the widest perimeter, suggesting that exposure to diverse swimming styles and conditions enhances the network's ability to generalise, even on specific test cases. In contrast, models with background subtraction (Models 5, 6, and 7) exhibit a noticeable contraction in the radar area. This suggests a counter-intuitive fact: removing the background simplifies the scene, but it likely introduces distortions, adds artefacts, or removes contextual water-turbulence cues that are actually useful for the network to localise submerged limbs.

This assessment is quantified by the Global Performance Index (GPI), reported in Table 1.

Table 1
Global performance index for the seven model configurations.

ID	GPI
Model 1	78.69
Model 2	68.20
Model 3	70.25
Model 4	72.58
Model 5	63.42
Model 6	65.76
Model 7	61.55

As detailed in Table 1, Model 1 achieves the highest GPI of 78.69, outperforming the specialised freestyle models (Models 2–4) and significantly surpassing the background-subtracted configurations (Models 5–7). Within the freestyle-only group, hyperparameters optimisation provides a tangible benefit (Model 4 reaches 72.58 vs. 70.25 of Model 3). However, the lower scores of Models 5–7 suggest that the applied background subtraction strategy may be too aggressive for the resolution of distal joints, reducing the overall sensitivity. These results highlight that data diversity (as in Model 1) plays a more critical role in underwater pose estimation robustness than dataset restriction or pre-processing aimed at static noise removal.

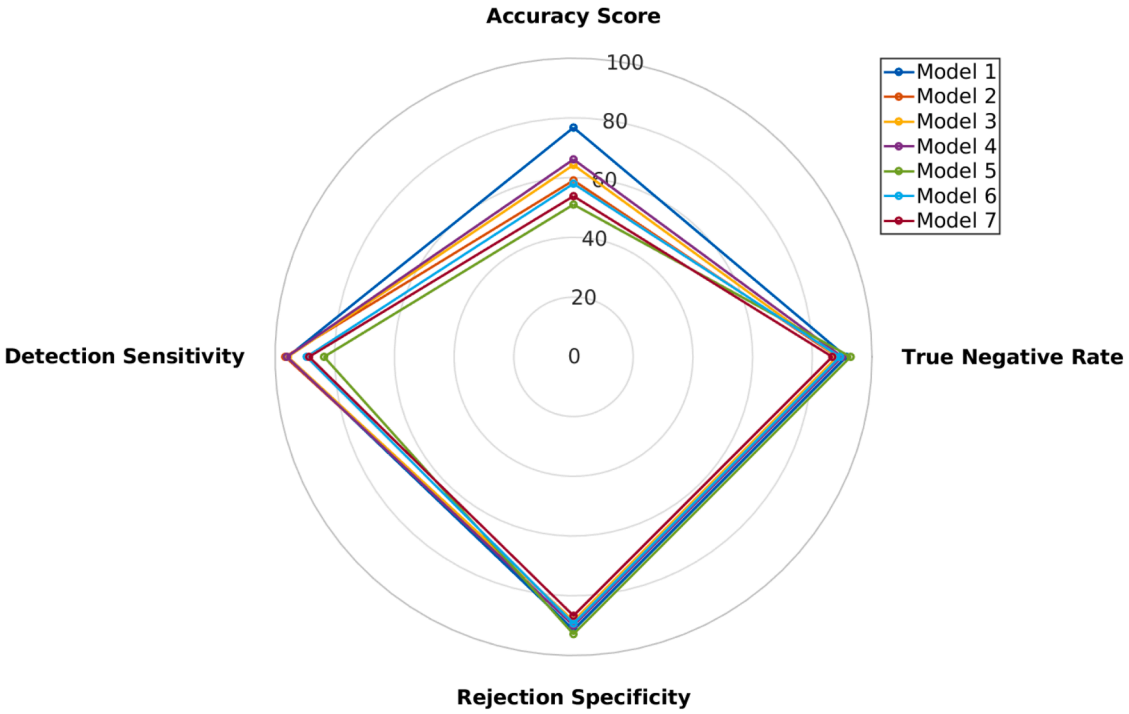

Fig. 6. Multi-model radar chart comparing the four aggregate performance indices.

Fig. 7 reports, for each model, the true score per-keypoint, defined as the percentage of ground-truth positive frames in which a keypoint is correctly detected within its tolerance radius (TP-in). The keypoints are ordered along the horizontal axis from the distal extremities (hand and foot tips) to the central segments (head and hips) and back to the distal joints on the contralateral side.

Across all models, a consistent pattern emerges. Central landmarks such as the head and hips exhibit the highest true scores, often approaching or reaching 100%, indicating a very reliable location whenever these points are visible. Intermediate shoulder, knee,and ankle segments generally achieve intermediate values, typically in the range 70–90%, with moderate variability between models. Distal extremities, especially the tips of the hand and foot, systematically show the lowest true scores, in several cases below 40%. This behaviour reflects the combined effect of smaller tolerance regions, higher motion dynamics, and more frequent extremity occlusions.

The differences between models are evident in both the absolute level and the stability of the curves. Early configurations trained on heterogeneous data (e.g. Model 1) achieve high performance on many central landmarks but exhibit pronounced drops at the extremities. Models incorporating background subtraction and regularisation (Models 6 and 7) tend to produce smoother profiles, with improved performance on several distal joints and less abrupt variations across the kinematic chain. These trends confirm that training set specialisation and regularisation strategies are critical for robust estimation of small, fast-moving anatomical landmarks.

To disentangle localisation accuracy from detection failures and spurious activations, the outcome statistics per-keypoint were summarised in the form of four heatmaps: NLA, FN rate, FP rate and TN rate (Figs. 8, 9, 10, 11). Each heatmap reports, for every model-keypoint pair, the corresponding conditional percentage, encoded by colour and annotated numerically.

The accuracy heatmap in Fig. 8 is derived from the normalised saturated error $e_{i,f}^{\text{sat}}$ and expresses, on a scale of 0 to 100, how precisely each landmark is localised when a prediction is available and the keypoint is visible. High values (yellow) correspond to small average errors relative to the tolerance radius, whereas low values (blue) indicate that the typical error is comparable to or larger than the admissible region.

The map confirms the core-periphery gradient already visible in the true-score curves 7. The head and hip keypoints are associated with very high accuracy for all models, while the tips of the hand and foot remain challenging, with scores often below 50. Intermediate joints show markedly model-dependent behaviour: some configurations obtain high accuracy on wrists and ankles, while others are less precise, suggesting that these joints are particularly sensitive to changes in training data and pre-processing.

Fig. 9 shows the false negative rate, calculated conditionally on ground-truth positive frames. This metric quantifies how often a visible landmark is not detected at all.

Central landmarks generally exhibit very low FN rates across all configurations, indicating that when the head or hips are visible, the networks almost always produce a prediction. In contrast, distal joints show substantially higher FN rates, with some model-keypoint combinations reaching values on the order of several tens of percent. One particular configuration with background subtraction without sufficient regularisation displays pronounced miss rates for certain keypoints (e.g. head and foot tips), suggesting that aggressive background processing can interfere with the network's ability to detect small or low-contrast structures. Regularised models reduce these miss rates, especially for problematic landmarks, while maintaining good performance on easier keypoints.

Figs. 10 and 11 report, respectively, the FP and TN rates conditioned on ground-truth negative frames. These metrics characterise how the models behave when a landmark is not annotated, e.g. because it is outside the field of view or completely occluded.

Early configurations trained on heterogeneous data show a marked tendency to produce spurious detections for the head on non-visible frames, with FP rates that can exceed 60–70% and correspondingly low TN rates. This behaviour suggests that the models rely strongly on contextual cues and may infer a head position even when the visual evidence is weak. As background subtraction and regularisation are introduced, head FP rates decrease dramatically and TN rates approach

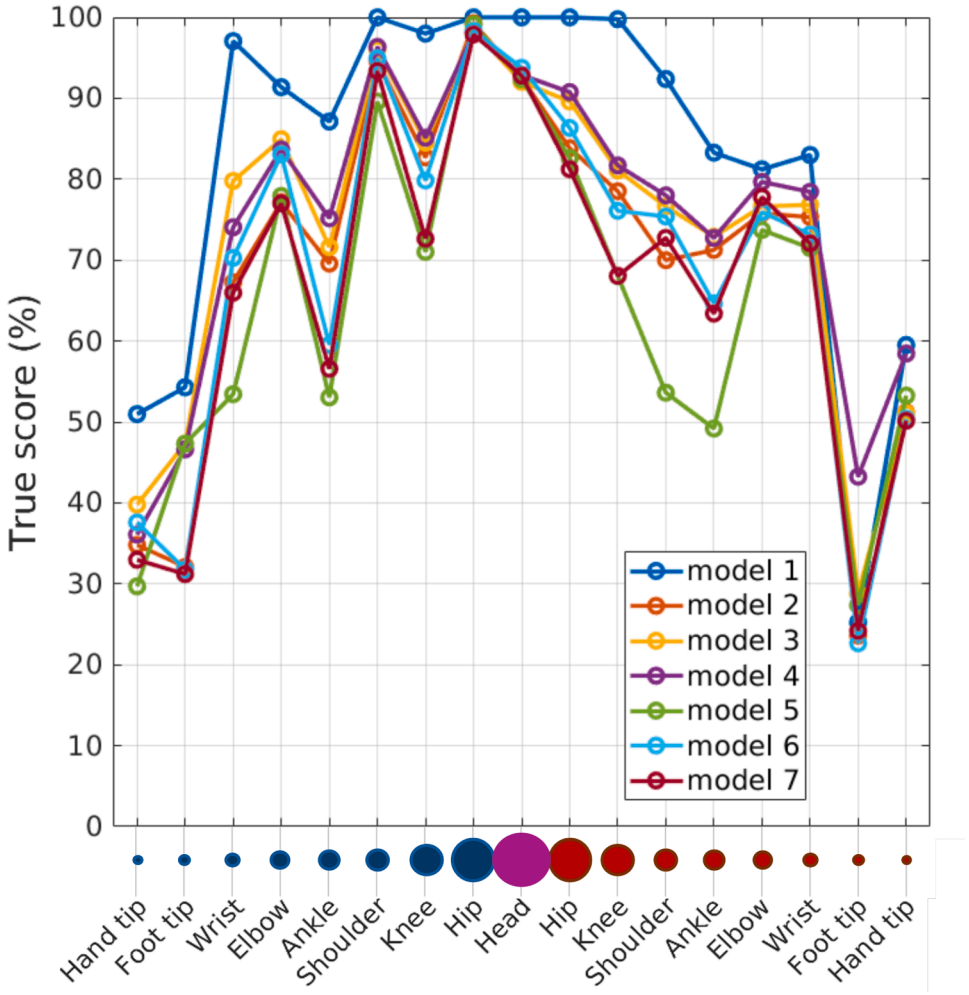

Fig. 7. Per-keypoint true score (TP-in rate on ground-truth positive frames) for all seven models. Each curve describes the fraction of correctly localized instances for each anatomical keypoint.

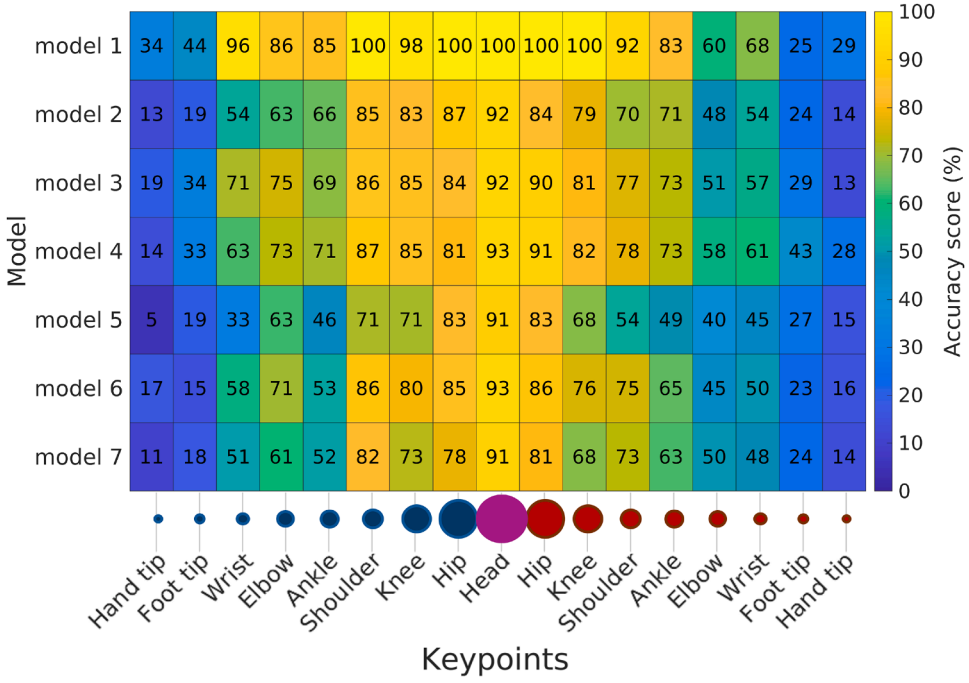

Fig. 8. Per-keypoint NLA for all models. Values close to 100 indicate that, when the keypoint is detected, its average distance from the ground truth is much smaller than the target-specific tolerance radius.

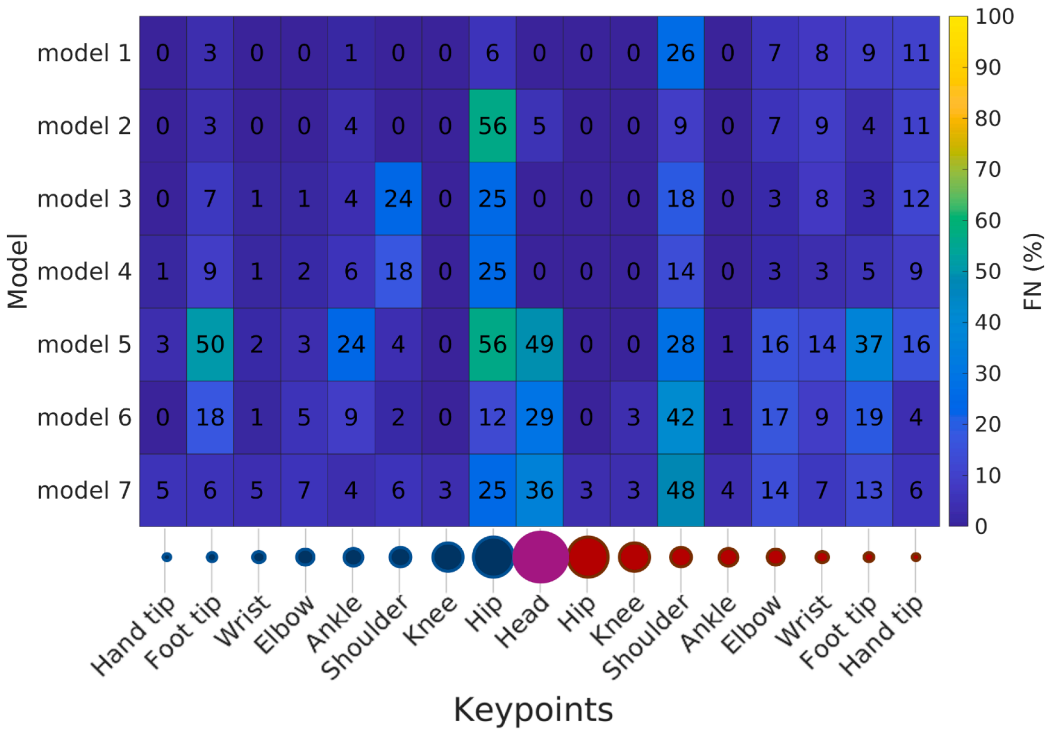

Fig. 9. False negative (FN) rate on ground-truth positive frames. Darker colors correspond to low miss rates, while lighter colors indicate that the model frequently fails to produce a prediction for a visible landmark.

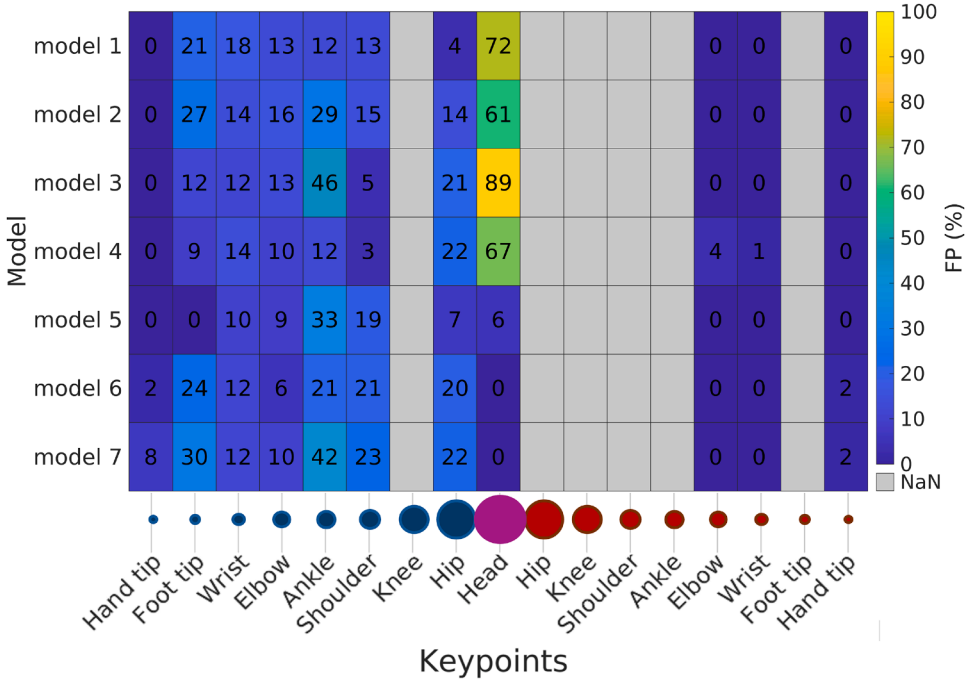

Fig. 10. False positive (FP) rate on ground-truth negative frames. High values indicate a tendency to hallucinate keypoints when they are not visible. Grey cells correspond to keypoints that are almost never labelled as non-visible in the test set.

100%, indicating a much more conservative and reliable visibility decision. For most other landmarks, the FP rates remain modest across models, although some configurations show elevated FP rates on specific joints (e.g. ankles), reflecting confusion with nearby structures or turbulent water patterns.

Together, the true score profiles and the visibility-conditioned heatmaps provide a detailed picture of the strengths and limitations of each configuration. All models achieve excellent localisation on central landmarks and satisfactory performance on intermediate joints, but distal extremities remain the main bottleneck, with higher miss rates and lower localisation accuracy. Training on stroke-specific datasets, combined with background subtraction and appropriate regularisation, leads to smoother per-keypoint performance profiles, reduced false positives on non-visible landmarks, and improved robustness on several

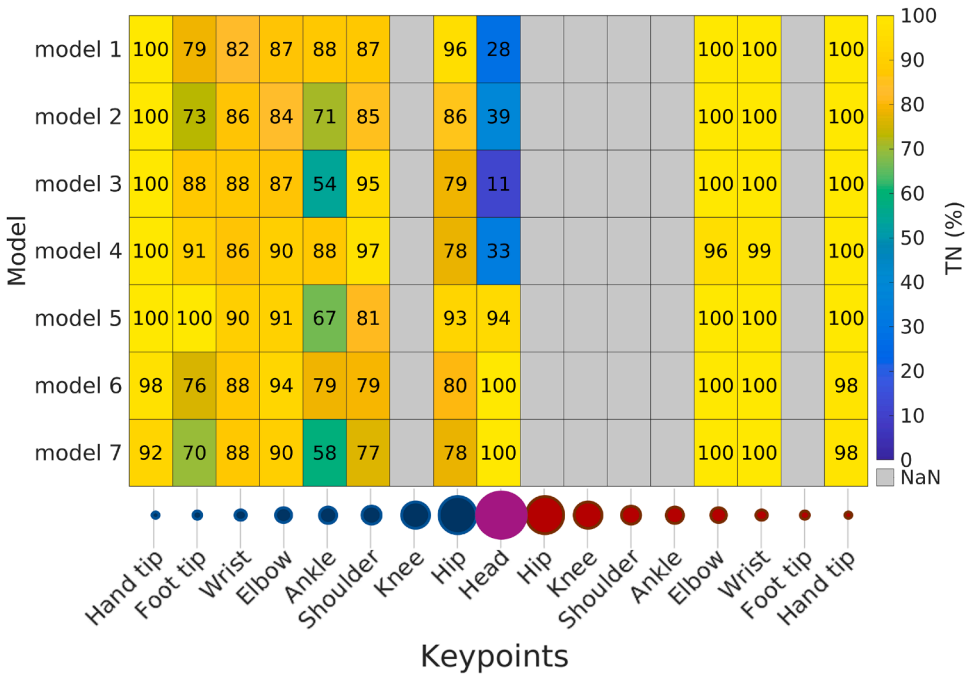

Fig. 11. True negative (TN) rate on ground-truth negative frames. High values indicate that the model correctly refrains from producing predictions when the keypoint is not annotated.

Table 2
Standard pose-estimation metrics.

Model	MPJPE [px]	PCK [%]	OKS
model 1	8.07	95.77	0.915
model 2	16.02	96.58	0.897
model 3	13.64	95.67	0.903
model 4	12.15	96.03	0.920
model 5	22.59	83.35	0.734
model 6	17.44	89.34	0.818
model 7	20.54	88.43	0.788

distal joints. These findings underline the importance of jointly considering localisation precision, sensitivity to visible keypoints, and specificity on non-visible ones when evaluating pose-estimation systems for underwater sports applications.

3.3. Comparison with standard pose-estimation metrics

For completeness, standard pose-estimation metrics were computed on the independent underwater test sequence used in this study. Table 2 reports the values of MPJPE, PCK and OKS for all model configurations.

4. Discussion and conclusions

The experimental results highlight a systematic anatomical performance gradient in the estimation of the pose of an underwater swimmer, with central body landmarks consistently outperforming distal extremities across all evaluated configurations. This behaviour reflects the intrinsic complexity of underwater imaging and the strongly multi-scale nature of the task, rather than being attributable to isolated modelling choices. Head and hip keypoints are generally detected and localised with high reliability, whereas hand and foot tips remain the most challenging landmarks, exhibiting lower detection rates and reduced localisation accuracy.

A central methodological outcome of this study is the clear distinction between binary validity metrics and continuous localisation quality measures. The percentage of True Positives within anatomical tolerance (TP_{in}) provides a strict assessment of compliance with anatomically meaningful error limits, but inherently discards information on the magnitude of localisation errors once the tolerance threshold is exceeded. This limitation becomes particularly evident for the distal extremities. Because their anatomical regions are small, the associated tolerance radius r_i is narrow and the binary decision boundary becomes extremely sharp. As a consequence, predictions that are only marginally outside the tolerance region are classified as complete failures, indistinguishable from gross localisation errors. This effect leads to an overly pessimistic evaluation of performance in small, fast-moving joints and masks the model's ability to approximate plausible trajectories.

The continuous NLA introduced in this work complements the binary metric by explicitly encoding the distance between the prediction and the ground truth, normalised by keypoint-specific tolerance. By retaining information on error magnitude, this score provides a more faithful representation of localisation quality, particularly in regimes where strict anatomical compliance is difficult to achieve. The combined use of TP_{in} and the continuous score therefore enables a clearer separation between detection reliability and spatial precision.

Moving beyond binary setting of thresholds, the multi-dimensional evaluation synthesises the trade-offs between different training configurations. The Global Performance Index identifies the heterogeneous training strategy (Model 1) as the most robust configuration. As shown by the radar representation, this model maintains the widest perimeter, indicating a balanced capability to preserve high sensitivity without sacrificing localisation accuracy. The analysis of per-keypoint heatmaps further shows that performance differences are not uniform across the kinematic chain, with central segments behaving consistently and distal extremities being more sensitive to training data composition and model configuration.

For completeness, the proposed evaluation was complemented with three standard pose-estimation metrics, MPJPE, PCK and OKS. These indicators show that most models achieve relatively high PCK values (above 95% for Models 0–3), OKS scores in the range 0.90–0.92, and MPJPE between approximately 8 px and 23 px, with the background-

subtracted configuration (Model 5) clearly underperforming across all three metrics. Overall, conventional measures suggest broadly comparable global accuracy for most configurations on this specific test sequence.

When compared with the tolerance-based analysis and the GPI framework (as shown in Table 1), this behaviour underlines some limitations of global metrics in the present context. MPJPE, PCK and OKS aggregate performance over all joints and frames and do not explicitly account for keypoint-specific scale or visibility, so they tend to mask the pronounced core-distal gradient and the different relevance of localisation errors across anatomical regions. The tolerance-normalised NLA and the associated per-keypoint indices are instead designed to incorporate these aspects explicitly and to separate localisation quality from visibility-related failures.

The observed core-periphery gradient has direct implications for biomechanical applications and analysis of sport performance. Central landmarks such as the head and hips are localised with high accuracy when detected, supporting a reliable reconstruction of trunk-level kinematics. Distal keypoints, in contrast, are affected by higher intrinsic uncertainty due to smaller tolerance regions, higher motion dynamics, and more frequent partial occlusions. From a biomechanical perspective, this suggests that kinematic variables derived from distal joints should be interpreted with greater caution and, where possible, complemented by temporal filtering or model-based constraints. Importantly, reduced binary validity at the extremities does not necessarily imply unusable predictions; rather, it often reflects predictions that are spatially close to the ground truth but fail to satisfy a strict anatomical threshold.

An additional contribution of this study lies in the explicit separation of detection outcomes conditioned on visibility of landmarks. By reporting False Negative and False Positive rates relative to ground-truth positive and negative frames, the analysis disentangles failures due to missed detections from spurious activations. This distinction is particularly relevant in underwater scenarios, where legitimate occlusions and out-of-field conditions are common and should not be conflated with localisation errors. Visibility-conditioned metrics reveal that high localisation accuracy on detected keypoints can coexist with substantial miss rates on difficult landmarks, reinforcing the need to jointly consider sensitivity, specificity, and spatial accuracy rather than relying on a single aggregate score. Furthermore, a basic intra-annotator repeatability test yielded a typical dispersion of only a few pixels, which is considerably smaller than the corresponding anatomical tolerance radius. This supports the use of manual labels as a metrological reference within the proposed evaluation framework.

The performance degradation observed in background-subtracted configurations can be attributed to the characteristics of the acquisition environment. The pool background is highly textured, including tiles, lane ropes, and light reflections, so the adopted medium-level background subtraction does not operate under the ideal assumption of a uniform static background. Under these conditions, the subtraction step can introduce artificial edges and discontinuities around the swimmer's silhouette, generating spurious features that interfere with keypoint localisation, particularly for small distal joints. Moreover, underwater turbulence and water texture patterns may provide contextual information that the network implicitly exploits during localisation; removing or distorting these cues can reduce the effective signal-to-noise ratio for limb detection. Therefore, the degradation observed in background-subtracted models in this study should not be interpreted as a general limitation of background removal, but rather as a consequence of applying a specific subtraction method in a non-uniform, feature-rich environment.

In conclusion, this work proposes a problem-orientated evaluation framework for 2D underwater swimmer pose estimation that explicitly accounts for anatomical scale and functional relevance. Keypoint-specific tolerances and continuous, tolerance-normalised accuracy scoring provide a more informative and reliable assessment than binary metrics alone in multi-scale scenarios. The results demonstrate that diversity of training data is a primary driver of robustness and generalisation, and that scale-aware evaluation is essential for correctly interpreting model performance in underwater sports biomechanics. Together, these findings support the adoption of evaluation protocols that are problem-related and metrologically grounded when deploying pose estimation systems for biomechanical analysis in challenging aquatic environments.

CRediT authorship contribution statement

Alessia Caputo: Writing – original draft, Validation, Supervision, Software, Methodology, Investigation, Data curation, Conceptualization; **Alberto Scocco:** Writing – review & editing, Visualization, Validation, Conceptualization; **Paolo Castellini:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Q. Dang, J. Yin, B. Wang, W. Zheng, Deep learning based 2D human pose estimation: a survey, Tsinghua Sci. Technol. 24 (6) (2019) 663–676. https://doi.org/10.26599/TST.2018.9010100
[2] S. Salisu, A.S.A. Mohamed, M.H. Jaafar, A.S.B. Pauzi, H.A. Younis, A survey on deep learning-Based 2D human pose estimation models, Comp., Mater. Continua 76 (2) (2023) 2385–2400. https://doi.org/10.32604/cmc.2023.035904
[3] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-Person 2D pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7291–7299. https://doi.org/10.1109/CVPR.2017.143
[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-Time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91
[5] Y. Xu, J. Zhang, Q. Zhang, D. Tao, ViTPose++: vision transformer for generic body pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. 46 (2) (2024) 1212–1230. https://doi.org/10.1109/TPAMI.2023.3330016
[6] H.M. Toussaint, M. Truijens, Biomechanical aspects of peak performance in human swimming, Anim. Biol. 55 (1) (2005) 17–40. https://doi.org/10.1163/1570756053276907
[7] M. Patil, R.H. Goudar, G.S. Hukkeri, AI For swimming recommendation systems: exploring the current landscape and research opportunities, Disc. Appl. Sci. 8 (2026) 156. https://doi.org/10.1007/s42452-025-08156-x
[8] P. Castellini, A. Scocco, A. Caputo, Measurement of the frontal area of a swimmer: alternative methods and uncertainty analysis, Acta IMEKO VN (VY) (2026) 1–9.
[9] A. Edriss, et al., Markerless motion analysis in biomechanics: current applications, challenges and future perspectives, Front. Physiol. (2025). https://doi.org/10.3389/fphys.2025.1649330
[10] C. Aulton, L. Wakili, B.W. Strafford, K. Davids, C.Y. Chiu, The application of deep learning human pose estimation in sport: a systematic review, Sport. Med. - Open 11 (155) (2025). https://doi.org/10.1186/s40798-025-00953-3
[11] M. Einfalt, D. Zecha, R. Lienhart, Activity-Conditioned continuous human pose estimation for performance analysis of athletes using the example of swimming, in: 2018IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 446–455. https://doi.org/10.1109/WACV.2018.00055
[12] X. Cao, W.Q. Yan, Pose estimation for swimmers in video surveillance, Multimed. Tools Appl. 83 (9) (2024) 26565–26580. https://doi.org/10.1007/s11042-023-16618-w
[13] N. Giulietti, A. Caputo, P. Chiariotti, P. Castellini, SwimmerNET: underwater 2D swimmer pose estimation exploiting fully convolutional neural networks, Sensors 23 (4) (2023) 2364. https://doi.org/10.3390/s23042364
[14] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D Human pose estimation: new benchmark and state of the art analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3686–3693. https://doi.org/10.1109/CVPR.2014.471
[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer, Cham, 2014, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[16] J.Y. Chiang, Y.-C. Chen, Underwater image enhancement by wavelength compensation and dehazing, IEEE Trans. Image Process. 21 (4) (2012) 1756–1769. https://doi.org/10.1109/TIP.2011.2179666
[17] M.J. Islam, Y. Xia, J. Sattar, Fast underwater image enhancement for improved visual perception, IEEE Rob. Autom. Lett. 5 (2) (2020) 3227–3234. https://doi.org/10.1109/LRA.2020.2974710
[18] Z. Kai, L. Jingyun, V.G. Luc, T. Radu, Designing a practical degradation model for deep blind image super-Resolution, Proc. IEEE Int. Confer. Comp. Vision (2021) 4771–4780. https://doi.org/10.1109/ICCV48922.2021.00475
[19] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, J. Wang, Human pose estimation using global and local normalization, 2017. arXiv:1709.07220
[20] C. Ge, W.F. Qin, ScaleFormer architecture for scale invariant human pose estimation with enhanced mixed features, Sci. Rep. 15 (1) (2025). https://doi.org/10.1038/s41598-025-12620-4
[21] R.B. Neupane, K. Li, T.F. Boka, A survey on deep 3D human pose estimation, Artif. Intell. Rev. 58 (1) (2024). https://doi.org/10.1007/s10462-024-11019-3