Engineering Applications of Artificial Intelligence 165 (2026) 113493

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Research paper

Automatic white shrimp (*Penaeus vannamei*) biometrical analysis from laboratory images using computer vision and deep learning

Remache González Abiam [a,*], Chagour Meriem [b], Bijan Rüth Timon [b], Trapiella Cañedo Raúl [b], Martínez Soler Marina [b], Lorenzo Felipe Álvaro [b], Shin Hyun-Suk [b], Zamorano Serrano María-Jesús [b], Torres Ricardo [d], Castillo Parra Juan-Antonio [e], Reyes Abad Eduardo [d], Ferrer Ballester Miguel-Ángel [c], Afonso López Juan-Manuel [b], Hernández Tejera Francisco-Mario [a], Penate-Sanchez Adrian [a]

[a] Institute SIANI, Universidad de Las Palmas de Gran Canaria (ULPGC), Las Palmas, 35001, Spain
[b] Aquaculture Research Group (GIA), Institute of Sustainable Aquaculture and Marine Ecosystems (IU-ECOAQUA), Universidad de Las Palmas de Gran Canaria (ULPGC), Telde, 35413, Spain
[c] Technological Centre for Innovation in Communications (iDeTIC), Universidad de Las Palmas de Gran Canaria (ULPGC), Las Palmas, 35017, Spain
[d] PRODUMAR S.A., Durán, 091650, Ecuador
[e] Biotechnology and Marine Genetic S.A. (BIOGEMAR S.A.), San Pablo, Santa Elena, 090350, Ecuador

ARTICLE INFO

Keywords:
Shrimp size estimation
Penaeus vannamei
Genetic assessment
Pose estimation
Computer vision
Deep learning

ABSTRACT

Manual morphological analysis for genetic selection in *Penaeus vannamei* aquaculture is a slow, error-prone bottleneck. We introduce Imashrimp, an automated system that uses colour and depth images to optimize this task by adapting deep learning and computer vision techniques to shrimp morphology. Imashrimp incorporates two discrimination modules to classify images by the point of view and determine rostrum integrity. These modules function as a "two-factor authentication" (human and Artificial Intelligence) system to validate annotations; this approach reduced metadata annotation errors, cutting point of view classification errors from 0.64% to 0% and rostrum integrity errors from 10.44% to 1.04%. A transformer-based pose estimation module predicts 23 keypoints on the shrimp's skeleton, achieving a general Mean Average Precision of 96.84% and a Percentage of Correct Keypoints of 91.67%. The resulting Two-Dimensional measurements are transformed into Three-Dimensional measurements using a Support Vector Machine regression. By achieving a final Mean Absolute Error (MAE) of 0.08±0.25 cm, IMASHRIMP demonstrates the potential to automate and accelerate shrimp morphological analysis, enhancing the efficiency of genetic selection and contributing to more sustainable aquaculture practices.

1. Introduction

Aquaculture provides a sustainable source of aquatic food that meets the nutritional demands of modern societies. In 2022, for the first time in history, aquaculture production surpassed that of capture fisheries, and it is expected to continue expanding in the coming years. This rapid expansion necessitates improvements in data collection and the development of novel analytical tools to ensure its sustainability. At the species level, white-leg shrimp (*Penaeus vannamei*) led global aquaculture production in 2022, with 6.8 million tonnes produced. Ecuador emerged as the world's leading exporter, in large part due to its sustained efforts to modernize production systems and implement genetic breeding programs (Food and Agriculture Organization of the United Nations (FAO), 2024).

In response to the anticipated increase in global shrimp demand, Ecuadorian shrimp farms must further enhance their competitiveness. A key strategy involves the adoption of non-invasive measurement methods in both production and breeding programs. These methods have been shown to reduce costs, enhance measurement efficiency, and improve final product quality (Ana et al., 2016).

The PMG-BIOGEMAR©genetic breeding program, developed by the University of Las Palmas de Gran Canaria (Spain), has been implemented by the Almar Group, a major shrimp producer based in Ecuador (Shin et al., 2020). Genetic selection is carried out using the Best Linear Unbiased Prediction (BLUP) methodology, in which thousands of shrimp are assessed for growth and morphological traits.

* Corresponding author.
E-mail address: abiam.remache101@alu.ulpgc.es (R.G. Abiam).

https://doi.org/10.1016/j.engappai.2025.113493
Received 13 September 2025; Received in revised form 2 December 2025; Accepted 5 December 2025
Available online 12 December 2025

Previous studies have established the genetic parameters of these morphological traits and identified the most relevant ones for selection in this population (Shin et al., 2023; Martínez Soler et al., 2024).

However, the trait measurements in these studies were obtained manually, a process that is time-consuming and prone to errors that are difficult to correct. Automating these measurements can significantly reduce operational costs and allow the evaluation of a much larger number of individuals, thereby strengthening the effectiveness of the genetic selection process. This study proposes a novel deep learning solution to automate these measurements, enabling precise and robust phenotyping.

This work introduces a system that uses deep learning to produce precise and robust shrimp measurements. We adapt the successful line of research pioneered by Wei et al. (2016) for human pose inference. Our system is designed to predict the coordinates of 23 keypoints in shrimp images. These keypoints collectively construct a virtual skeletal structure that encapsulates the morphological characteristics of interest. By mapping these points, the system effectively represents the anatomical framework of the shrimp, allowing the extraction and analysis of relevant morphological variables.

Our approach estimates the specific keypoints that define the start and end of each required measurement. For example, measuring the head length first requires estimating the animal's virtual skeleton; the Two-Dimensional (2D) distance is then calculated between the keypoints at the tip and back of the head. Once the 2D keypoints have been detected and correctly estimated, we perform a regression that transforms that 2D measurement to the required Three-Dimensional (3D) measurement. Our system is able to perform both the required lateral and dorsal measurements.

Our method's goal is to enhance the manual measurement process performed by human operators. The manual process of measuring and labelling shrimp is prone to human error. In this paper, we focus on using artificial intelligence (AI) as a "second check" to validate human annotations and raise an alarm if discrepancies are found, although our results show the AI already commits fewer errors than human operators. Furthermore, our approach enhances operational efficiency by drastically reducing the annotation time per specimen from approximately 9 min (manual measurement of all traits) to just 32 ms (automated inference per image), directly translating into significant labour cost savings and increased throughput for breeding facilities.

Our approach not only measures keypoints robustly but also handles human error in two additional ways: by detecting the shrimp's point of view (lateral or dorsal) and by checking its *rostrum integrity*. The *rostrum* is the beak-shaped structure on the shrimp's head, which is essential for measuring traits like cephalothorax length (Martínez Soler et al., 2024). As this structure is often broken during the shrimp's life, *rostrum integrity* simply refers to the assessment of whether this 'horn' is intact (good) or broken, a critical factor that determines if certain keypoints can be measured.

First, an AI detection system predicts the shrimp's point of view (lateral or dorsal). This works as a secondary system that detects when the human might have produced an error and generates an alert. Second, a similar AI detector validates the human's assessment of the rostrum's integrity. If our AI detector predicts that the human could have committed an error (e.g., attempting to measure a broken rostrum as if it were intact (good)), we raise an alarm. In summary, the main contributions of our work are the following:

- **An automatic and robust system that can measure shrimp.** We introduce a system for robust 3D shrimp measurement, achieved by modelling the animal as a 23-keypoint skeleton and using deep learning to extract multi-point-of-view morphological data from over 12,000 annotated images.
- **Two AI detection systems capable of mitigating human error.** We deploy two complementary AI modules that validate the animal's point of view (lateral/dorsal) and rostrum integrity, alerting the operator to potential data entry errors.
- **A regression estimator to increase the precision of the 3D measurements.** We use a regression estimator to convert 2D pixel-based keypoints into real-world 3D measurements, which significantly reduces the final system error.

2. Related work

We will now detail the context in which our work resides. We will first detail other works performed in the field of aquaculture engineering that address the extraction of morphological/biometrical information from animals in a wide array of contexts. Secondly, we will describe the work performed in the field of articulated 3D pose estimation, which is the line of work from which our 3D shrimp pose estimation draws inspiration.

To the best of our knowledge, no one has performed 3D shrimp pose estimation before, the closest work was that of Chirdchoo et al. (2024) that only performed visual analysis and reports errors of 2.1 centimetres (cm) of Mean Absolute Error (MAE) for the length of the shrimp. In our work we are capable of estimating not just the length of the shrimp but also a complete set of morphological measurements from our detected shrimp skeleton. When comparing only the length estimation we provide a much more precise measurement yielding 0.54 cm of MAE, which improves the results by nearly an order of magnitude. This clearly shows that a more in-depth assessment of morphological traits, like the one we propose in this paper, greatly improves overall quality.

2.1. Computer vision in morphological analysis

The morphological measurement of aquatic animals is a critical aspect of fisheries management, species monitoring, and aquaculture. Although advanced artificial intelligence techniques and automated systems have been widely applied to fish species, similar approaches for shrimp remain under-explored. The challenge of automated morphological analysis is twofold: first, the extraction of variables in pixel space, and second, the conversion of those pixels into real-world physical measurements (e.g., centimetres).

For the pixel extraction task, methods range from classical Computer Vision to advanced deep learning, including semantic segmentation, detection, and pose estimation. For the conversion task, methods vary from using no conversion to simple scaling factors or regression models. Our work argues that for complex genetic selection, a pose estimation framework combined with a regression-based converter is the superior approach.

Existing approaches to pixel extraction vary in complexity. Early work on shrimp relied on traditional Computer Vision (Harbitz, 2007; Hadiyanto and Widodo, 2022), while recent studies use Convolutional Neural Networks (CNNs) (Chao Zhou and Yang, 2021). However, these modern systems, for both shrimp and fish, often rely on segmentation methods (e.g.,Mask Region-based Convolutional Neural Network (Mask R-CNN) Chirdchoo et al., 2024; Zhou et al., 2023; García-Santamaría et al., 2022; Garcia et al., 2019; Huang et al., 2020, You Only Look Once (YOLO) Climent-Perez et al., 2024; Dong et al., 2023; Tonachella et al., 2022).

While effective for simple external metrics like total length (often the only variable measured), these methods are fundamentally insufficient for genetic selection. Our objective is a complete morphological analysis of 23 distinct variables, many of which depend on internal anatomical keypoints (like segment junctions) that segmentation masks cannot locate. For instance, identifying the precise positions from where to measure abdominal segments requires detecting subtle anatomical landmarks rather than just the animal's outline; a segmentation mask might accurately capture the shrimp's silhouette but fail to pinpoint these internal articulation points required for specific genetic traits (e.g., segment lengths), leading to measurement inaccuracies due to boundary misalignment.

Table 1
Comparative analysis of Imashrimp with state-of-the-art morphological analysis systems. Abbreviations: N.° vars.: Number of measured variables; TL: Total Length; MAPE: Mean Absolute Percentage Error; SLCNet: Shrimp Larvae Counting Network; YOLACT: You Only Look At CoefficienTs.

Research	Species	Objective	Method	Dataset	N.° vars.	TL MAE (cm)	TL MAPE (%)
Chirdchoo et al. (2024)	Pacific white shrimp	Estimating body weight by extracting five key morphological features: area, perimeter, width, length, and body posture	Detectron2 (Wu et al., 2019a) + Classical Computer Vision	Train: 3946; Test: 1036	5	2.10 (Known scaling factor)	14.57 (Known scaling factor)
Harbitz (2007)	*Pandalus borealis*	Automatically estimate the length of the shrimp shell	Classical Computer Vision	Not given	1	Non-comparable metrics	Non-comparable metrics
Hadiyanto and Widodo (2022)	*Penaeus vannamei*	Estimation of body weight using morphometric features extracted from images	Classical Computer Vision	Train: 20; Test: 6	1	Non-comparable metrics	Non-comparable metrics
Chao Zhou and Yang (2021)	*Cherax quadricarinatus* (Shrimp larvae)	Estimation of body length measurement	SLCNet (Liu et al., 2022)	Train: 294; Test: 126	1	No conversion to physical units	No conversion to physical units
Zhou et al. (2023)	Shrimp, species not given	Automatically estimate the size (length and width) of shrimps to monitor their growth rate.	Mask RCNN (He et al., 2017) + Classical Computer Vision	Train: 300; Test: 150	2	No conversion to physical units	No conversion to physical units
Climent-Perez et al. (2024)	12 species + 1 due to sexual dimorphism	Estimating the length of fish	YOLACT++ (Bolya et al., 2019)	Train: 1108; Test: 152	1	1.76 (Visual metrology based on homography)	11.44 (Visual metrology based on homography)
García-Santamaría et al. (2022)	Lampuga (Dolphinfish), *Coryphaena hippurus*	Estimate the average fork length of fish in each landing box	Mask R-CNN (He et al., 2017)	Train: 246; Test: 30	1	Not given	4.00-6.90 (Length-weight scale factor)
Garcia et al. (2019)	7 species of pelagic fish	Measurements of the length of individual fish	Classical Computer Vision + Mask R-CNN (He et al., 2017)	Train: 1625; Test: 80	1	No conversion to physical units	No conversion to physical units
Huang et al. (2020)	Fish (species not specified)	Measurement of body dimensions (length and width) of fish in an unrestricted environment	Mask R-CNN (He et al., 2017) + Classical Computer Vision	Not given	2	0.55 (Scale factor with chessboard pattern)	4.00 (Scale factor with chessboard pattern)
Voskakis et al. (2021)	*Gilthead seabream* and *European seabass*	The distance between the mouth and tail, mouth and eye, and eye and tail is estimated	Open Pose (Cao et al., 2021)	Train: 250; Test: 20	3	Not given	Seabream: 3.15; Seabass: 7.40 (Scale factor with chessboard pattern)
Dong et al. (2023)	Fish (species not specified)	Detection of 7 biological keypoints on the fish's body	YOLO (Redmon et al., 2016) + Lite-HRNet (Yu et al., 2021) (7 keypoints)	Detection: 3000; Pose Estimation: 2000	Not given	No conversion to physical units	No conversion to physical units
Tonachella et al. (2022)	*Gilthead seabream*	Automatic estimation of body length and prediction of weight	YOLO (Redmon et al., 2016) + ResNet-101 (He et al., 2016)	Detection: 1400; Pose Estimation: 12800	2	1.15 (Scale factor with chessboard pattern)	5.50 (Scale factor with chessboard pattern)
Imashrimp (Ours)	*Penaeus vannamei*	Complete morphological analysis for genetic selection	Imashrimp framework	Train: 11122; Test: 1245	23, see Fig. 6	0.54 (Support Vector Machine)	3.76 (Support Vector Machine)

This limitation led us to a pose estimation framework. While some fish studies have used pose estimation (e.g., OpenPose Cao et al., 2021, Lite-High-Resolution Network (HRNet) Dong et al., 2023 or Residual Network 101 (ResNet-101) Tonachella et al., 2022) to extract more variables (2 to 3), Imashrimp is the first to apply a state-of-the-art Vision Transformer (ViT) architecture (Xu et al., 2022) to shrimp.

This architectural choice aligns with recent trends in complex agricultural and biological visual analysis, where transformer-based (ViT) models are increasingly favoured over traditional Convolutional Neural Networks (CNNs) for their superior ability to capture global relationships, despite potential trade-offs in processing speed (Çakmak, 2025). Our novel, reusable framework (integrating discrimination, pose estimation, and conversion) is purpose-built for the genetic selection workflow. It leverages a large, high-quality dataset (12,367 annotated images) to achieve high-precision results across all 23 keypoints, not just total length.

In Table 1, we show an in-depth comparison between Imashrimp and other approaches that perform morphological analysis on other fish or crustacean species. This table highlights the methodology, the number of variables measured (N.° vars.), and the reported error for Total Length (TL), the most common benchmark. As shown in Table 1, our framework not only measures a far more comprehensive set of variables (23) but also achieves a significantly lower error (3.76% Mean Absolute Percentage Error (MAPE)) than the most directly comparable study (Chirdchoo et al., 2024) on the same species (14.57% MAPE).

Regarding pixel-to-centimetre conversion modules, there are several methodologies. Some studies, such as (Chao Zhou and Yang, 2021), Zhou et al. (2023), Garcia et al. (2019) and Dong et al. (2023), do not perform physical measurements, instead prioritizing the detection of precise morphological variables. Other works, such as (Huang et al., 2020; Voskakis et al., 2021) and Tonachella et al. (2022), employ triangulation systems using a chessboard and binocular cameras for calibration.

In addition, some studies use a known real-world measurement to derive a scaling factor, as seen in Chirdchoo et al. (2024), García-Santamaría et al. (2022) and Climent-Perez et al. (2024). Finally, regression models provide another approach for efficient pixel-to-centimetre conversion. Our study opts for this methodology, given the substantial amount of real data collected for the 23 morphological variables, allowing for the development of a robust predictive model for this task.

2.2. Articulated 3D pose estimation

Obtaining the pose of a human or an animal has been extensively researched for years and great advances have been made. The pose of humans or animals is in essence an articulated skeleton, and the task lies in finding with precision the 3D locations of the joints of said skeletons.

Before deep learning, the most promising results obtained in 2D human pose estimation can be seen in Andriluka et al. (2014). Obtaining precise 3D measurements remained a challenge at that time. With the appearance of Convolutional Pose Machines (CPM) by Wei et al. (2016) it was showed that using deep learning could yield very robust 3D estimations to the articulated pose estimation problem. This advancement was made possible by creating datasets that addressed the need for larger amounts of training data, some of these datasets were Human 3.6, by Ionescu et al. (2014) and HumanEva, by Sigal et al. (2010).

From the original CPM research paper, many others continued to improve the 3D estimation. For example, in the work by Tome et al. (2017) it is proposed to optimize the 2D and 3D positions together to improve both tasks through the inherent sharing of information within the neural network architecture. In the work by Moreno-Noguer (2016) the 3D pose is obtained by modelling the problem as a regression between two Euclidean distance matrices. In subsequent years, the paradigm changed from using convolutional neurons to using visual transformers by Dosovitskiy et al. (2021). This paradigm change was then incorporated into many human pose estimation approaches like (Dosovitskiy et al., 2021) and Xu et al. (2022).

Concerning the application of said approaches to animal pose estimation, there has been less work overall, but the said techniques have been demonstrated to be robust enough to handle dog poses, like (Rueegg et al., 2022), or even zebra, tiger, elephant, and horse, as seen in Yao et al. (2022) amongst others. Again, this has been possible due to the creation of datasets like the ones by Xu et al. (2023) and by Marshall et al. (2021). The majority of the articulated 3D animal pose estimation has been focused on mammals, which makes our work quite unique as it shows that such techniques can be used on a wider array of animal species, and particularly those of great economical interest.

3. Background

In this section, we will describe the scenario in which our system works to facilitate the understanding of our paper. Firstly, we will introduce the way in which genetic selection is performed, secondly, we will show the image capture setup (in Fig. 1) we have used for our data collection and that is also used for testing, and finally, we will describe the keypoint virtual skeleton definition that we propose and that comes directly from the initial and final position from which measurements are performed by the expert geneticists. **Shrimp selective breeding.** Genetic selection breeding programs for the species allow breeders to be selected according to their Estimated Breeding Values (EBVs) for a desired trait to obtain the next generation. For the current population, weight and morphological traits are among the most important traits to be selected. To perform such selection, a statistical analysis of morphological traits using Best Linear Unbiased Prediction (BLUP) is used. Some of the more costly morphological traits to obtain are the precise measurements of each part of the shrimp. Our system uses 23 morphological measurements from both the dorsal and lateral point of views of the animal.

A detailed description diagram of the morphological measurements, and how they are related to the shrimp virtual skeleton, can be found in Fig. 6. More qualitative examples of the morphological measurements as they are performed in real cases can be seen in Fig. 2 for the lateral case and in Fig. 3 for the dorsal case. Fig. 2 provides a complete visual breakdown for the lateral point of view, illustrating the comparison between the ground truth and the model's prediction (top row), the pixel-level error and activation heatmaps (middle row), and a diagram of the derived morphological measurements (bottom row).

Shrimp virtual skeleton definition. As explained before, each of the measurements done on the shrimp for selective breeding consists of a starting and final point of measurement. Many of those points are used several times. For example, when measuring the head of the shrimp we measure from point x_1 till point x_2, and afterwards when measuring the length of the first segment of the body of the shrimp we measure from point x_2 to point x_3. By taking each of the points required for measurement and their topology we have our proposed virtual skeleton as seen in Fig. 2 (top row on the left).

In our work we estimate two virtual skeletons depending on the point of view, the lateral skeleton, Fig. 2, and the dorsal skeleton, Fig. 3. Given the skeletons that we have defined, we can perform keypoint pose estimation, similar to the one used in humans in Wei et al. (2016) or (Xu et al., 2022), to learn to predict the points we require for our measurements. We learn two separate neural networks that estimate the lateral and the dorsal skeleton separately.

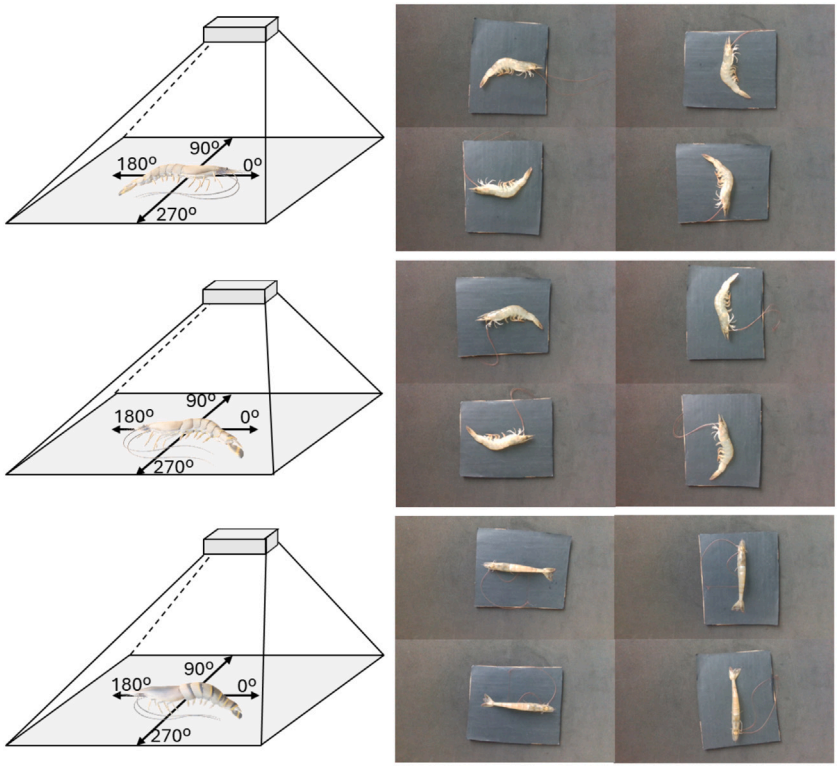

Fig. 1. Image acquisition setup and methodology. All shrimp were captured in four configurations, 0°, 90°, 180° and 270°. **First row:** Image acquisition of the shrimp's right lateral point of view from all degrees. **Second row:** Image acquisition of the shrimp's left lateral point of view at all degrees. **Third row:** Capture of images of the shrimp's dorsal point of view at all degrees. This consists of a total of 12 images captured for each shrimp specimen, with a total number of 1223 shrimp specimens in the annotated dataset.

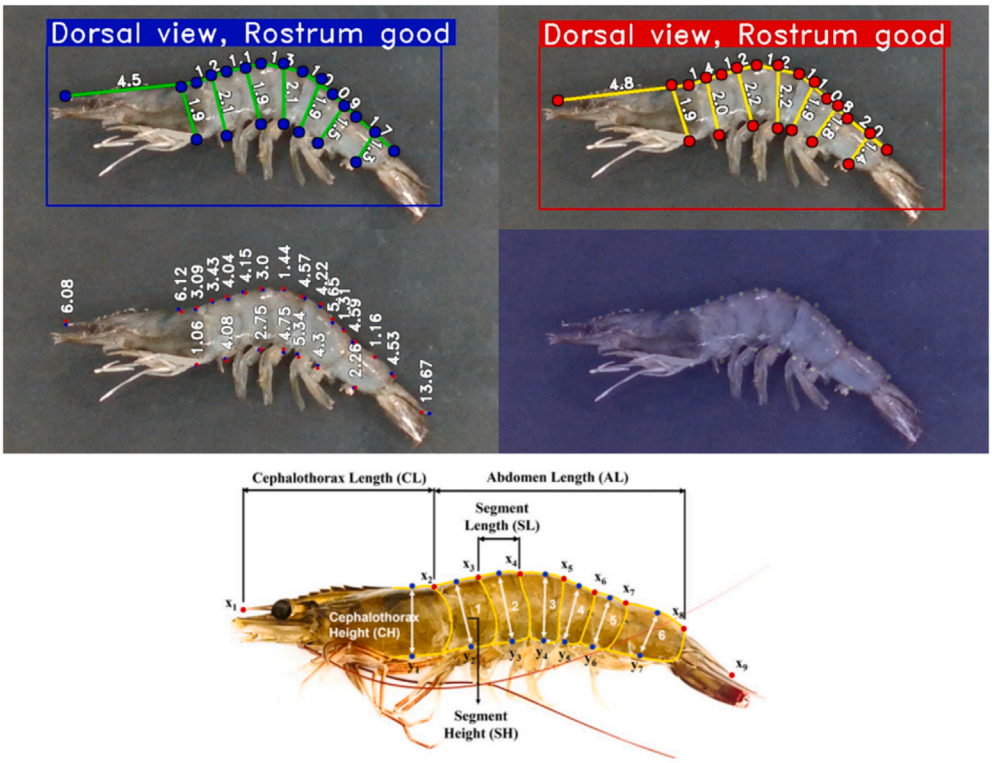

Fig. 2. Description of the keypoint virtual skeleton used by our shrimp pose estimator and the measurements performed on the animal used for genetic improvement. Top row: Ground truth virtual skeleton (left) and the prediction by Imashrimp (right), both with measures in centimetres. **Middle row:** Pixel errors (left) and heatmaps activations (right). **Bottom row:** Derived measurements diagram. A full description of all measurements can be found in Shin et al. (2023) and in Martínez Soler et al. (2024).

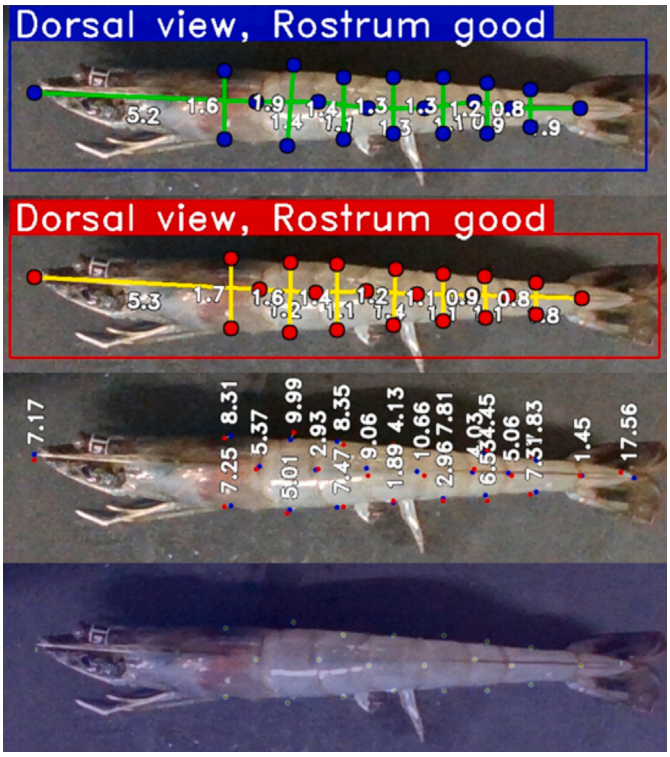

Fig. 3. Description of the dorsal keypoint virtual skeleton used by our shrimp pose estimator. First row: Ground truth virtual skeleton, with measures in centimetres. **Second row:** Prediction of Imashrimp with measures in centimetres. **Third row:** Pixel errors. **Fourth row:** Heatmaps activations.

4. Method

We have named our approach Imashrimp, and it is composed of three different elements: firstly, two modules based on artificial intelligence that perform the discriminator tasks (point of view and *rostrum* integrity) to avoid human error during data labelling, secondly, the shrimp pose estimation system that performs the keypoint estimation from which measurements can be performed, and lastly, a regression module that learns how to convert from pixel coordinates into real world positions from which the 3D measurements can be obtained.

The function of the discriminator modules for the discrimination of images will be explained according to two main factors, the shrimp's point of view and the shrimp integrity of the *rostrum*. Afterwards, the operation of the shrimp pose estimation module will be explained, which will be responsible for detecting 23 keypoints for each of the point of views (lateral and dorsal), if the *rostrum* is broken it will only detect 22. Finally, the morphological regression module will be explained, which is responsible for converting the morphological variables resulting from the detection of keypoints (pose estimation system) from pixels to centimetres.

The proposed method integrates these three modules (discrimination, pose estimation, and regression) into a complete system. To illustrate this, the system's end-to-end workflow is presented in Fig. 4. This flowchart details the logical process from image capture, through the logical decisions of the discrimination modules, to the final data generation, clarifying the workflow for research and industrial use. Complementing this workflow, Fig. 5 details the technical architecture of the core pose estimation module. The specific morphological variables derived from this process are then shown in Fig. 6.

4.1. Discrimination systems

As described above, our proposed approach incorporates two independent neural networks that are used to detect the conditions of the shrimp alongside human annotation to greatly reduce human error. When humans create the metadata associated with the images, they have to introduce if the image was taken from a lateral or dorsal point of view and if the shrimp has a complete *rostrum*. Based on this information, our system selects the specific pose estimation model that is required, whether the shrimp requires lateral measures or dorsal measures to be detected, or reduce the number of keypoints from 23 to 22 if the rostrum is not present.

To maximize the robustness of our system, we show that the best approach is to use both the human annotation and the automatically detected artificial intelligence (AI) results. This works as a sort of two-factor authentication, if human and AI agree, the data is introduced into the database, if they disagree an alarm is raised for the data to be checked and corrected. With this scheme, we manage to reduce human error from 0.64% to 0% for the annotation of the point of view of the image (lateral/dorsal), and we reduce human error in *rostrum* presence from 10.44% to 1.04% with our discriminator systems. Results can be seen in Table 2.

Both our **rostrum integrity** and **point-of-view** classifiers work in the same way to help the human technician. To classify between the lateral/dorsal point of view and the presence of rostrum, we use a Residual Network 50 (ResNet-50) architecture, by He et al. (2016), for binary classification. The use of Convolutional Neural Networks (CNNs) to classify images as a proxy for complex physical measurements is an analogous task to recent work in other fields, such as medical diagnostics, where numerous Deep Learning architectures have been benchmarked for classifying intraoral photographs to predict cephalometric measurements (Kartbak et al., 2025). While our binary task did not require such an extensive benchmark, this prior work validates our general approach.

We pass the complete image to train the classifier to detect the desired prediction, as seen in Fig. 1. The human makes an assessment Ψ^{h}_{pov} for the point of view of the shrimp (lateral or dorsal) and an assessment for the *rostrum integrity* Ψ^{h}_{ri}. We define these outputs as binary, e.g., $\Psi_{pov} \in \{0,1\}$, where 0=Lateral and 1=Dorsal and $\Psi_{ri} \in \{0,1\}$, where 0=Broken and 1=Good. In parallel, our discriminator architecture makes a parallel assessment Ψ^{AI}_{pov} and Ψ^{AI}_{ri}. The final decision to raise an alert, $A \in \{0,1\}$ (where 1=Alert), is made as follows:

$$A_{pov} = \Psi^{AI}_{pov} \oplus \Psi^{h}_{pov} \tag{1}$$

$$A_{ri} = \Psi^{AI}_{ri} \oplus \Psi^{h}_{ri} \tag{2}$$

Using the exclusive OR (XOR) operator, denoted by $\oplus$, this formulation ensures that an alert ($A = 1$) is raised only if the human and AI assessments disagree (i.e., one is 0 and the other is 1). This does not fix the very few cases in which both commit a mistake, this in any case is a lesser problem as it does not introduce errors that in the previous approach were not present already.

We believe that our approach provides the best of both worlds, human errors occur in repetitive tasks from humans losing focus, whereas AI errors occur due to different factors. Due to our hybrid two-factor authentication, the AI can fix the most human errors due to loosing concentration and humans can avoid errors that the AI might introduce.

4.2. Shrimp pose estimation

We have modelled the task of measuring the shrimp morphology as that of a shrimp pose estimation task, where each of the joints of the skeleton are the keypoints that will be used to estimate the desired measurements. We draw from the rich state-of-the-art advances in human pose estimation, in which the best performing approaches

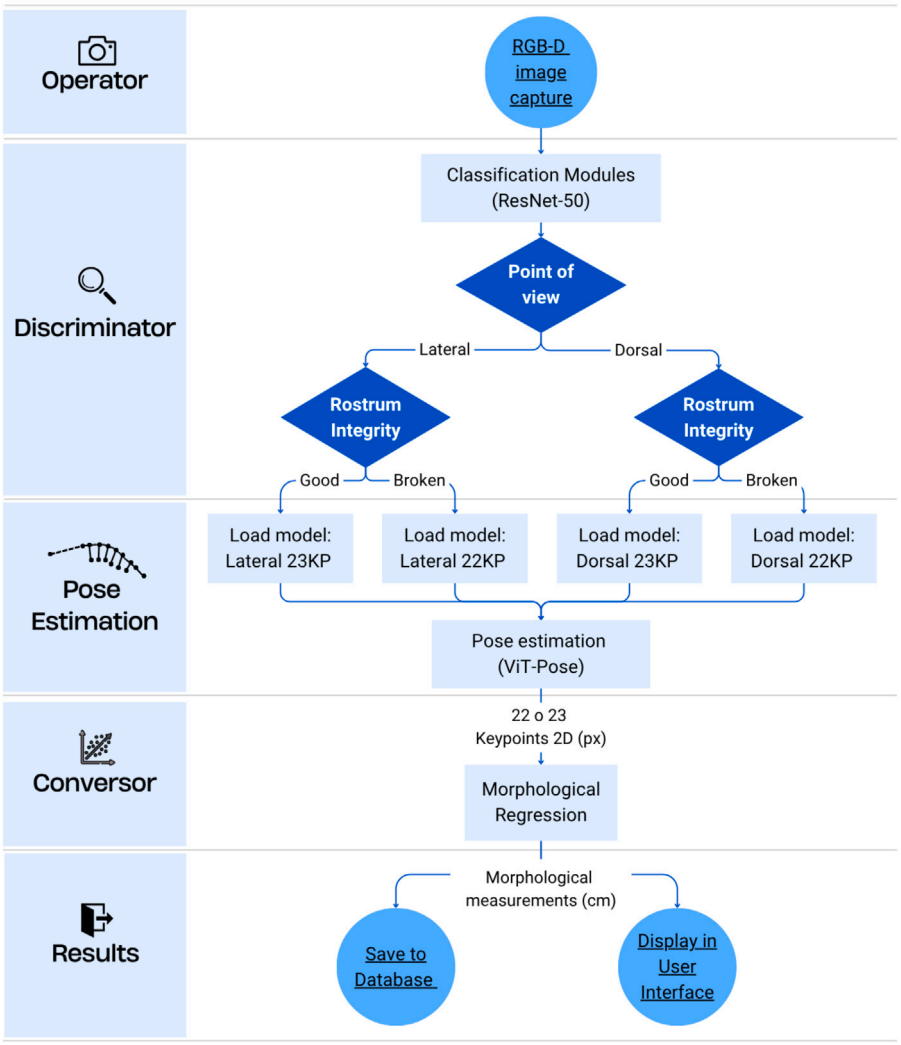

Fig. 4. The end-to-end operational workflow of the Imashrimp system. Abbreviations: ViT-Pose: ViT-based pose estimation networks; KP: Keypoints.

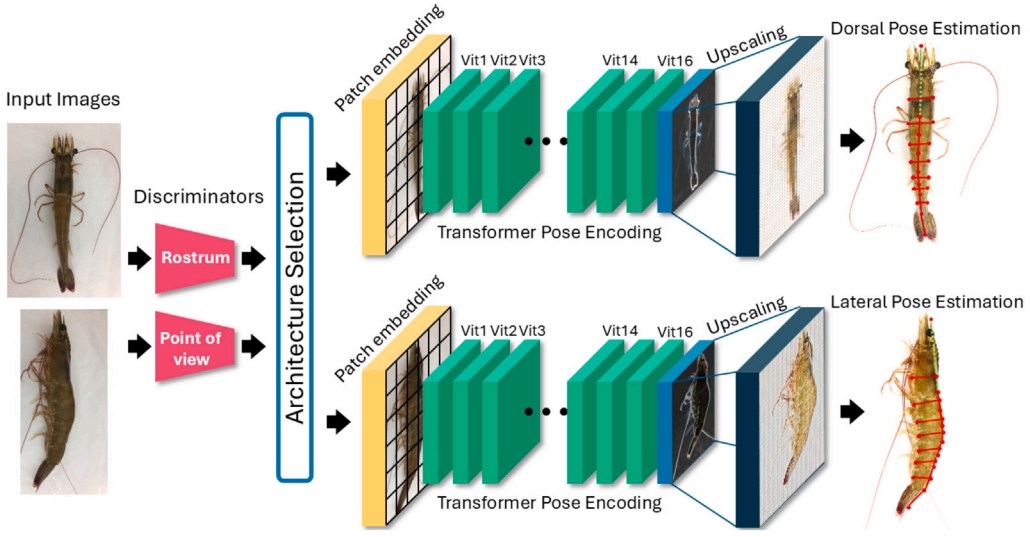

Fig. 5. Technical overview of the proposed method for shrimp pose estimation and size regression. The diagram illustrates the flow from input images through the discriminators to the parallel ViT-based pose estimation networks (ViT-Pose). The two architectures shown (Dorsal and Lateral) represent four distinct models (dorsal-22, dorsal-23, lateral-22, and lateral-23) by adapting the skeleton from 23 keypoints (rostrum good) to 22 (rostrum broken), indicated by the green dotted line in the final estimation.

use a neural network to learn to predict a series of heatmaps from which the keypoint locations are derived. We follow the work of Xu et al. (2022) for our neural network by creating an encoder/decoder architecture. Encoding is done using the Vision Transformer (ViT) architecture by Dosovitskiy et al. (2021). The decoder consists of a bilinear layer followed by a $ReLU$ activation function and a final pose predictor as described in Xiao et al. (2018). In our work we employ Red-Green-Blue-Depth (RGB-D) images to further enhance the precision of our keypoint estimations given the 3D information that the depth channel provides.

Given an input RGB-D image of size $X \in R^{HxWx4}$, where H is of 192 pixels of height and W is 256 pixels of width, we perform an initial encoding in a patch embedding space F_0 of smaller resolution. Our embedding reduces the resolution by a factor of $d = 16$ and has dimensions of $C = 1280$, which creates a patch embedding of size $F_0 \in R^{\frac{H}{d}x\frac{W}{d}xC}$ which in our work leaves us with $F_0 \in R^{12x16x1280}$. From this initial encoding F_0 we apply several ViT layers which consist of a multi-head self-attention layer Θ_i and then a multi-layer perceptron MLP_i. Layer normalization is applied before every F_i block, we represent it by the $\widehat{*}$ symbol.

We use 16 ViT layers for our encoder. The dimensionality of the embedding C remains unchanged throughout the encoding. The final form of the encoding F_i at each intermediate step i is defined by the standard Transformer block architecture, which consists of a Multi-Head Self-Attention (MSA) layer (Θ) followed by a Multi-Layer Perceptron (MLP), both with residual connections:

$$F_i' = F_{i-1} + \Theta_i(\widehat{F_{i-1}}) \tag{3}$$

$$F_i = F_i' + MLP_i(\widehat{F_i'}) \tag{4}$$

Our decoder architecture is a simple combination of a bilinear interpolation and a ReLU activation function, with the final pose predictor. Given the final encoding output of our 16 ViT layers $F_{16} \in R^{\frac{H}{16}x\frac{W}{16}xC}$, the decoder creates a set of heatmaps per each keypoint k of the virtual skeleton and upscales by a factor of 4. The number of keypoints in our configuration is of $N_k = 23$ for the general case and of $N_k = 22$ for animals without *rostrum*. This yields the tensor $F_{heat} \in R^{\frac{H}{4}x\frac{W}{4}xN_k}$. Given F_{heat}, the predictor optimizes an L_2 loss from the training data to learn the prediction of the final position of each keypoint k. We show in Fig. 5 a visual diagram of the shrimp pose estimation network.

In order to train such a neural network, up to 3.6 million images can be required, as seen in Ionescu et al. (2014). To avoid such high costs in annotation (our dataset is 12367 images), we leverage weights trained on different tasks to bootstrap our training through transfer learning. Our encoder has been pre-trained in other tasks/datasets to have better encodings that will allow the pose estimation to be successful. The datasets on which our network has been pre-trained are the Microsoft Common Objects in Context (MS-COCO) dataset, by Lin et al. (2015), the artificial intelligence (AI) challenger dataset, by Wu et al. (2019b) and the Max Planck Institut Informatik (MPII) Human pose dataset, by Andriluka et al. (2014).

4.3. Shrimp morphological regression

Once the keypoints have been found in the 2D image space, the next step is to derive the real 3D distances. In order to do so, the simple way to approach it is by calibrating how many centimetres does one pixel equate to, this was the method used by the IMAFISH method, by Navarro et al. (2016). We consider this approach our baseline method. However, as described by Garcia et al. (2019) it yields much more precise measurements to construct a regression model from known instances. Following their discoveries, we construct a regression model per measurement.

In the general case, our pose estimator returns a set of 23 keypoints X_i and from those 23 points a set of 22 measurements D_a. We refer again to Fig. 6 for details of each keypoint and their related measurements. The desired measurements must be in 3D, which we define as D_a^{3D}, but the ones we obtain from our shrimp keypoint detection network are in 2D, defined on the image plane, which we define as D_a^{2D}.

To obtain the 3D measurements from the 2D, we propose to learn a regression model. If one seeks to find a mapping between our 2D measurements D_a^{2D} and the real 3D measurements D_a^{3D} it can be posed as a problem of learning the coefficients of the function $D_a^{3D} = D_a^{2D} * \alpha + \beta$. Due to the inaccuracies of our data, which come from the human measurements, pixel quantization effects and the camera parameters, an exact solution of α, β does not exist. Due to this reason regression tries to find the closest hyperplane, as close to flat as possible, that models the relation between our 2D and 3D measurements. It does so by performing the optimization described by Vapnik et al. (1996). Support Vector Machine (SVM) regression is well-suited for this task because, due to the noise in the measurements (e.g., keypoint jitter), the mapping from 2D pixel distances to 3D measurements is no longer perfectly linear. The epsilon-insensitive loss function of SVM allows the model to ignore small errors within a defined margin, providing robustness against this noise and ensuring better generalization than standard least-squares methods. We use this regression model as it proved to be the best performing approximation. We obtain an individual regression per measurement, the regression function is learned from the training samples of D_a^{3D}, D_a^{2D} pairs. If we compare our regression approach with the baseline of just performing a simple calibration, we can see in Table 6 that there is a substantial advantage to perform regression.

5. Experimental setup

This section will present the different experiments that were carried out to validate the proposed approach. First, the dataset used in the experiments is described. Then, each module, i.e., the Point of View Discriminator, the Rostrum Discriminator, the Shrimp Pose Estimation and the Shrimp Size Regressor will be validated separately, each with a set of experiments aimed at demonstrating the performance of the different modules. Finally, an overall validation will be conducted for the whole proposed system.

5.1. Dataset

The images used in this article correspond to the SABE (Servicio de Análisis para Acuicultura *y* Biotecnología de Alta Especialización) laboratory and were captured during an eight month period (August 2023 to April 2024). The images were captured using an Intel Realsense D435 RGB-D camera. The camera was fixed with a tripod in a zenithal position at 30 centimetres of the plane in a controlled laboratory environment featuring a black background and similar uniform lighting conditions.

The dataset is comprised of 12 images per animal specimen, with each consisting of an Red-Green-Blue (RGB) image and its corresponding depth map. The same animal is photographed from three different point of views, lateral right, lateral left, and dorsal point of view at four different angles; see Fig. 1. To clarify how these point of views were handled, the 'lateral right' and 'lateral left' images were not used to train separate networks. Instead, they were combined into a single 'lateral' dataset to train both the 'lateral-22' and 'lateral-23' models, leveraging the anatomical symmetry of the keypoint skeleton from either point of view.

After the imaging procedure, 2856 images containing rotten shrimp and 2621 images that were blurred were manually discarded. The final resulting dataset contains 12 367 shrimp images, with a total of 1223 individuals photographed. Regarding rostrum integrity, the dataset exhibits a significant class imbalance, comprising 10,764 images with intact ('good') rostrums and 1603 with broken rostrums (see Table A.11). To mitigate the classification bias towards the majority

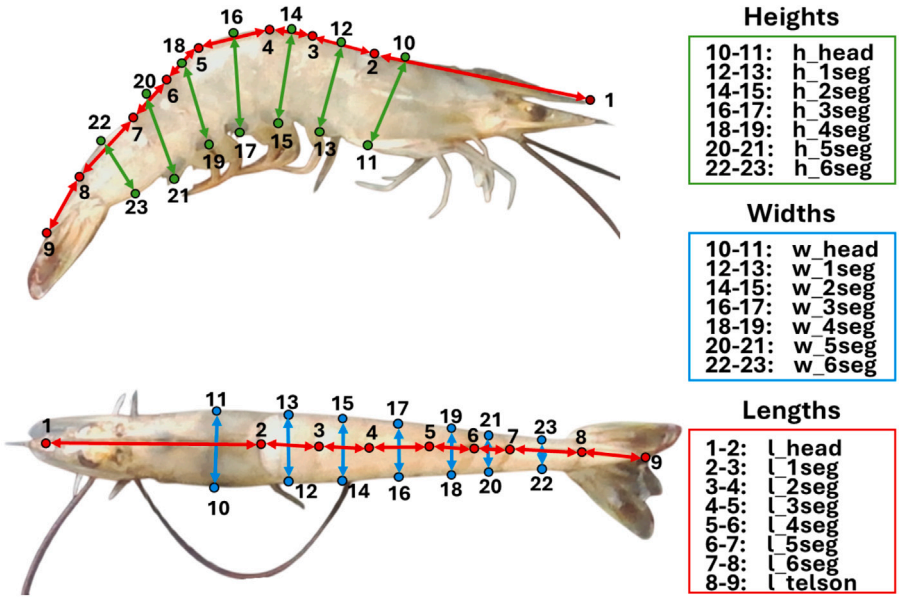

Fig. 6. Description of the keypoint virtual skeletons (lateral and dorsal) and the extracted morphological measurements. First row: Shrimp lateral keypoint virtual skeleton, keypoints 1 to 9 can be identified (red points), representing morphological variables of length. Keypoints 10 to 23 can be identified (green points), representing morphological variables of heights. **Second row:** Shrimp dorsal keypoint virtual skeleton, keypoints 1 to 9 can be identified (red points), representing morphological variables of length. Keypoints 10 to 23 can be identified (blue points), representing morphological variables of widths.

class inherent in such imbalances, we implemented the "two-factor" (Human-AI) authentication system described in Section 4.1. This collaborative approach serves as a critical quality control layer, correcting potential AI misclassifications driven by the data disparity. For testing purposes 10% of all shrimp specimens, and all their images, were separated to ensure a robust validation, all shrimp specimens used for testing were not seen during training. We include a complete description on the dataset creation and details in Appendix A.

For the discriminator modules, ground truth data was created to ensure a robust verification mechanism. During image capture, researchers concurrently recorded information about the point of view (Lateral or Dorsal) and the rostrum's integrity (Good or Broken). This dataset includes annotations from 12 367 images, comprising both human observations and those designated as ground truth.

The ground truth data was created by having an expert geneticist check the work of the technicians that performed the data annotation to record the cases of human error. The dual-phase documentation approach aims to enhance data reliability by comparing real-time observations, as done by technicians on-site, with curated post-acquisition assessments.

All 12 367 images in the dataset were annotated with the 23 keypoints that form the virtual skeleton for the 'Shrimp Pose Estimation' module. Then, using the estimated skeleton keypoints, we extract the morphological variables of length, height, and width. Examples of ground truth labels and morphological variables can be found in Fig. 6. The annotations were created using a generic annotation tool, CVAT (Computer Vision Annotation Tool), and subsequently exported in the MS-COCO keypoint format Lin et al. (2015).

For the Shrimp Size Regressor, during image capture, information was also collected about the actual morphological variables of the animal, identical to those shown in Fig. 6. This information was used in the regression model.

5.2. Proposed discriminators experiments

These study employs a ResNet-50 (He et al., 2016) inspired neural network to improve the assessment of shrimp morphology through binary discriminators of two different key features: point of view and rostrum integrity. The model's performance will be evaluated by measuring error rates (Error %) on a test set, comparing errors from: (1) human researchers, (2) the discriminator, and (3) their combined system.

To quantify the improvement offered by the hybrid system, the trained model will be compared the case in which only human assessment and the case in which only artificial intelligence (AI) assessment is provided. The idea is to demonstrate improvements in accuracy and synergy between human expertise and machine learning.

5.3. Proposed pose estimation experiments

The experiments conducted for our Pose Estimation modules aim to evaluate the performance in the test image subset and demonstrate its utility for the proposed task. The test subset has been made by choosing shrimp specimens randomly from the whole image set and making all images of that specimen to be part of the test set. By doing this we assure that all test individuals have never been seen during training.

The input images are first processed through the dual discriminator system, which assesses the point of view and rostrum integrity. Based on the initial classification, the images are routed to one of four specific estimation neural networks: **1** (Lateral point of view + 23 keypoints), **2** (Dorsal point of view + 23 keypoints), **3** (Lateral point of view + 22 keypoints), **4** (Dorsal point of view + 22 keypoints).

The four pose estimation neural networks are trained independently using type-specific images with the keypoint skeleton annotations adapted accordingly. The pose estimation networks will be tested individually, to asses the precision of the 2D estimations, and altogether to asses the precision of the final 3D measurements after the whole process is performed.

We will use commonly used metrics such as Euclidean Pixel Error (EPE) (Rong and Gang, 2024), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) per identified keypoint of the overall system, i.e. the four modules working as a complete system. For individual modules, we evaluate performance using: (1) Mean Average Precision (mAP) as in Xu et al. (2022), and (2) Percentage of Correct keypoints (PCK) (Andriluka et al., 2014), which applies a pixel threshold to determine correct keypoint detection.

5.4. Proposed size regression experiments

The Support Vector Machine (SVM) regression model was selected to accurately convert the 23 shrimp morphological variables, as illustrated in Fig. 6, from pixels to centimetres. To achieve an efficient system for converting pixel measurements to centimetre measurements, two approaches are compared. First, a non-regression method is implemented using a scale factor derived from ruler images, calculated based on the pixel distance between the 0 cm and 1 cm marks. And secondly, the one we chose, an SVM regression-based method that uses real measurements of all morphological variables in both centimetres (from the image capture phase) and pixels (from annotated ground truth data) to learn the regression coefficients.

The objective is to identify the conversion approach that delivers the most accurate conversion from pixels to centimetres. Width and height measurements can only be achieved through an specific point of view as shown in Fig. 6. But length measurements can be acquired through both point of views. We will show a comparison of precision between using either dorsal or lateral measurements to obtain length measurements.

6. Results

This section presents results from both quantitative and qualitative perspectives for all experiments: the Discriminator modules, Pose Estimation module, and Shrimp Size Regressor module. To validate the entire system, we will use a test subset with 1245 images, approximately 10% of the total annotated dataset with 12367 images.

6.1. Discriminator results

In this section we will describe the results we obtained when applying both of our artificial intelligence (AI) based discrimination modules to reduce human annotation error. Implementation details will also be described for all experiments. For all binary classification tasks in the discriminator modules, a standard confidence threshold of 0.5 was applied to the model's output probabilities to determine the predicted class labels.

6.1.1. Point of view discriminator results

Our point of view discriminator was trained to classify shrimp images into lateral and dorsal point of views using a dataset of 12367 annotated images. The network was trained for five epochs with a learning rate of 0.002 and a batch size of 256.

Using the test subset, we evaluated error rate, correct and incorrect detections across three classifiers: (1) human experts, (2) artificial intelligence (AI), and (3) our proposed hybrid system. Comparative performance results are presented in Table 2. Of the whole test image set human researchers made 8 errors (0.64%). The discriminator module produced 0 errors (0%).

Finally, testing the "Human-AI (Ours)" system on the 1245 images yielded a 100% accuracy rate (0.0% error), correctly classifying all 831 lateral and 414 dorsal images. This result indicates that the AI module, in this controlled test set, effectively eliminates the 0.64% error rate introduced by human-only annotation.

6.1.2. Rostrum integrity discriminator results

The proposed rostrum discriminator was trained to classify shrimp images based on rostrum integrity (good or broken), using the same dataset as the previous discriminator. The network was trained for five epochs with a learning rate of 0.0005 and a batch size of 256.

Using the test subset, we evaluated the error rate, correctly classified and incorrect detections with three classifiers: (1) human experts, (2) artificial intelligence (AI), and (3) our proposed hybrid system. The comparative results are shown in Table 2. Of the whole test image set human researchers made 130 errors (10.44%). The discriminator module produced 37 errors (2.97%).

Finally, testing the complete system on the 1245 images achieved a much improved error rate of (1.04%). In this case the hybrid approach that combines AI and human assessments is by far the best performing approach.

6.2. Pose estimation results

The pose estimation system includes four independent neural network architectures that are used based on the animal's point of view (lateral or dorsal) and the rostrum integrity (good or broken). The discriminators separate the 1245 test images into four groups, assigning each to an appropriate pose estimation module.

To validate our choice of a ViT-Pose architecture, we conducted a comparative analysis against two other state-of-the-art pose estimation baselines: YOLO-Pose (Jocher and Qiu, 2025) and High-Resolution Network (HRNet) (Yu et al., 2021). The performance of our four specialized Imashrimp modules, alongside the comparative results from the other baselines, is summarized in Table 3. The results show that our specialized, ViT-Pose-based system (Imashrimp) achieves the best general performance on the 1245 test images, with a mAP of 96.84% and a Percentage of Correct Keypoints at a threshold of 10 pixels (PCK@10px) of 91.67%, which justifies our architectural choice.

The 2D keypoint errors were estimated by comparing the pixel predictions from the test set against the annotated ground truth. A summary of this analysis is presented in Table 4, which aggregates the results by the two main virtual skeletons (Lateral and Dorsal), as they represent distinct anatomical structures. A more granular analysis is provided in Appendix B, which stratifies these 2D errors across all four specialized sub-models and includes the 'Score' (heatmap confidence) metric.

We show in Fig. 7 the heatmaps generated by the pose estimation module. They have been increased by a factor of four to provide clear visibility, the original heatmaps can be seen in Fig. 3 and in Fig. 2.

It can be seen that the activation peaks are sharply concentrated over the specific anatomical joints of interest, demonstrating that the model is not relying on irrelevant background features. This high-precision activation gives us confidence that the model is learning the correct underlying representation of the shrimp's skeleton as intended.

6.3. Size regressor results

This section details the conversion of 2D keypoints into 3D real-world measurements (cm). We first conduct an extensive benchmark to select the optimal regression model (SVM), and then we evaluate its final performance against a baseline scale-factor method and alternative data views.

To determine the most robust regressor, we first conducted an extensive benchmark of 14 different regression models. The performance of these models, evaluated on their ability to convert 2D pixel measurements to 3D real-world measurements (cm), is detailed in Table 5. To clarify, the 'Sig. (vs. SVM)' column indicates the statistical significance of each method's MAE when compared to the 'SVM (Baseline)' method using a paired Wilcoxon signed-rank test.

While several models (e.g., K-Neighbors and Polynomial regression) yielded a Mean Absolute Error (MAE) comparable to the Support Vector Machine (SVM), the SVM was selected as the optimal model. This decision was based on its low MAE combined with a more constrained 95% Confidence Interval (CI), which indicates a higher level of prediction, consistency, and reliability.

Having justified the selection of SVM, we then performed a high-level comparison between the optimized regression-based approach (SVM) and the baseline scale-factor (non-regression) method. This comparison is shown in Table 6.

Table 2
Classification performance of human, artificial intelligence (AI), and hybrid approaches for point of view and rostrum integrity discrimination.

Point of view discriminator				Rostrum integrity discriminator			
Classifier	Error (%)	Correct	Incorrect	Classifier	Error (%)	Correct	Incorrect
Human	0.64	1237	8	Human	10.44	1115	130
AI	0.00	1245	0	AI	2.97	1208	37
Human-AI (Ours)	**0.00**	**1245**	**0**	Human-AI (Ours)	**1.04**	**1232**	**13**

Table 3
Performance of the pose estimation modules on the test dataset (D_{test}images) depending on the classification of the discriminator modules for Imashrimp, Yolo and HRNet.

		mAP 50:95 (%)			PCK@10px (%)		
Sub-Model	N° images	Imashrimp (Ours)	Yolo	HRNet	Imashrimp (Ours)	Yolo	HRNet
dorsal-22	73	96.44	83.73	94.86	93.28	88.48	90.90
lateral-22	144	98.40	85.86	95.24	90.18	81.00	88.54
dorsal-23	341	93.11	71.50	89.89	90.86	80.24	85.43
lateral-23	687	99.40	88.97	95.97	92.38	80.32	82.96
General	**1245**	**96.84**	**82.51**	**93.82**	**91.67**	**82.51**	**86.99**

Table 4
Comparative 2D error analysis between real and predicted keypoints for the Lateral and Dorsal point of view. Abbreviation: px: pixel.

Point	Lateral Point of View			Dorsal Point of View		
	EPE (px)	RMSE (px)	MAPE (%)	EPE (px)	RMSE (px)	MAPE (%)
1	11.97 ± 36.29	27.02	0.92	12.54 ± 15.64	14.17	1.22
2	3.19 ± 1.78	2.58	0.30	3.98 ± 2.26	3.24	0.37
3	3.42 ± 2.07	2.83	0.32	3.91 ± 2.31	3.21	0.35
4	4.59 ± 3.45	4.06	0.42	4.46 ± 2.83	3.73	0.40
5	4.90 ± 3.27	4.17	0.46	5.20 ± 3.30	4.35	0.46
6	3.62 ± 2.18	2.99	0.35	3.78 ± 2.09	3.05	0.34
7	3.31 ± 1.94	2.71	0.31	3.65 ± 2.06	2.96	0.34
8	3.64 ± 2.15	2.99	0.36	3.62 ± 1.95	2.91	0.37
9	8.79 ± 5.14	7.20	0.71	9.78 ± 5.33	7.87	0.89
10	5.58 ± 7.24	6.47	0.49	6.33 ± 5.05	5.73	0.55
11	7.36 ± 8.53	7.97	0.68	6.55 ± 5.44	6.02	0.58
12	3.52 ± 2.17	2.92	0.33	4.39 ± 2.60	3.60	0.40
13	6.09 ± 5.23	5.68	0.54	4.49 ± 2.66	3.69	0.41
14	3.88 ± 4.28	4.09	0.35	3.95 ± 2.40	3.27	0.36
15	4.69 ± 4.59	4.64	0.42	4.08 ± 2.41	3.35	0.37
16	4.42 ± 3.75	4.10	0.41	4.11 ± 2.44	3.38	0.37
17	4.54 ± 3.22	3.94	0.41	4.30 ± 2.54	3.53	0.39
18	4.12 ± 5.28	4.74	0.40	4.25 ± 2.60	3.52	0.39
19	4.75 ± 5.33	5.05	0.44	4.07 ± 2.39	3.33	0.37
20	3.42 ± 2.02	2.81	0.33	3.98 ± 2.22	3.22	0.38
21	4.96 ± 3.18	4.17	0.47	3.86 ± 2.24	3.15	0.35
22	3.88 ± 2.42	3.24	0.39	4.65 ± 2.69	3.80	0.45
23	3.76 ± 2.22	3.09	0.36	4.69 ± 2.66	3.82	0.46
General	**4.80 ± 8.23**	**6.75**	**0.44**	**4.92 ± 4.64**	**4.78**	**0.45**

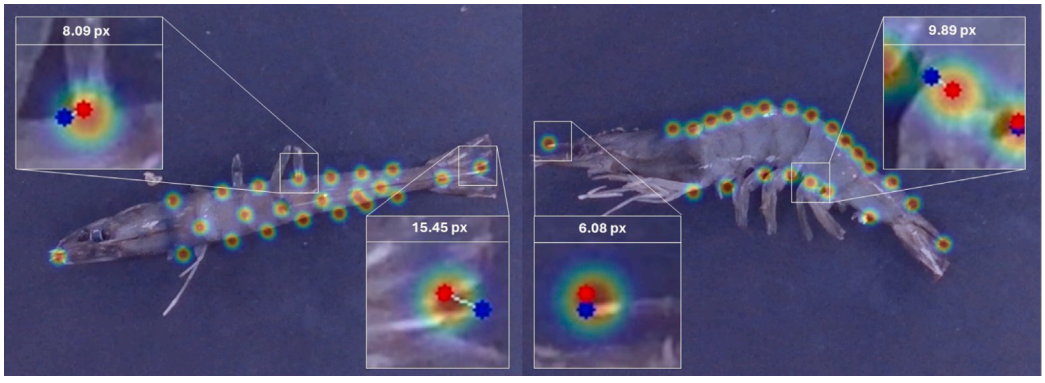

Fig. 7. Visualization of pose estimation interpretability and keypoint accuracy. The model's activation heatmaps are overlaid on representative dorsal (left) and lateral (right) test images. Magnified insets illustrate the Euclidean Pixel Error (EPE), showing the distance between the model's prediction (red point) and the ground truth (blue point).

Table 5
Comparative performance analysis of regression methods for 2D keypoint to 3D measurement conversion. All models were evaluated using 2D pose estimations generated by the baseline ViT-Pose Huge configuration. Abbreviations: IC% (MAE): 95% Confidence Interval, SGB: Stochastic Gradient Boosting.

Method	MAE (cm)	IC 95% (MAE)	RMSE (cm)	MAPE (%)	*p*-value (vs. SVM)
Ridge	0.10 ± 0.30	[0.09, 0.11]	0.32	5.14	4.0e−2
SGB	0.09 ± 0.27	[0.08, 0.10]	0.28	4.79	3.0e−2
Quantile	0.10 ± 0.30	[0.09, 0.11]	0.32	5.02	1.5e−2
MLP	0.14 ± 0.51	[0.13, 0.15]	0.49	5.60	1.0e−4
Linear	0.12 ± 0.35	[0.11, 0.13]	0.37	5.44	9.0e−3
Log-Lin	0.15 ± 0.47	[0.13, 0.16]	0.48	5.78	4.0e−4
Lin-Log	0.10 ± 0.30	[0.09, 0.11]	0.31	5.13	2.5e−2
Log-Log	0.12 ± 0.38	[0.11, 0.14]	0.40	5.46	5.0e−3
Polynomial	0.08 ± 0.26	[0.07, 0.09]	0.27	4.64	8.3e−2
k-neighbors	0.08 ± 0.26	[0.07, 0.09]	0.27	4.67	6.0e−2
Kernel	0.11 ± 0.34	[0.10, 0.12]	0.44	5.73	2.0e−3
Decision Tree	0.09 ± 0.27	[0.08, 0.10]	0.30	4.95	3.1e−2
Random Forest	0.08 ± 0.27	[0.08, 0.09]	0.29	4.83	1.5e−1
SVM	**0.08 ± 0.25**	[0.07, 0.08]	**0.25**	**4.56**	-

Table 6
Comparative analysis of the 3D error between ground truth and Imashrimp's predicted measurements for all white shrimp morphological variables according to the conversion method. (Significance test: *** $p < 0.001$).

Conversion not using regression					Conversion using SVM regression				
Variable	MAE (cm)	IC 95% (MAE)	RMSE (cm)	MAPE (%)	Variable	MAE (cm)	IC 95% (MAE)	RMSE (cm)	MAPE (%)
total	0.92 ± 1.06	[0.75, 1.10]	1.31	6.55	total	0.54 ± 0.69	[0.45, 0.62]	0.77	3.76
abdomen	0.22 ± 0.23	[0.75, 1.10]	0.29	2.31	abdomen	0.20 ± 0.23	[0.17, 0.24]	0.26	2.03
l_1seg	0.10 ± 0.12	[0.09, 0.13]	0.14	8.00	l_1seg	0.09 ± 0.11	[0.08, 0.10]	0.11	6.76
l_2seg	0.12 ± 0.11	[0.09, 0.12]	0.15	10.79	l_2seg	0.07 ± 0.09	[0.06, 0.08]	0.09	6.18
l_3seg	0.10 ± 0.14	[0.09, 0.12]	0.15	7.89	l_3seg	0.09 ± 0.12	[0.07, 0.10]	0.12	6.29
l_4seg	0.16 ± 0.11	[0.16, 0.19]	0.18	16.47	l_4seg	0.09 ± 0.11	[0.08, 0.10]	0.11	8.34
l_5seg	0.22 ± 1.09	[0.10, 0.34]	1.09	11.30	l_5seg	0.20 ± 1.10	[0.05, 0.30]	1.11	7.79
l_6seg	0.12 ± 0.10	[0.12, 0.15]	0.15	6.21	l_6seg	0.06 ± 0.08	[0.06, 0.07]	0.08	3.25
l_head	0.30 ± 0.38	[0.25, 0.34]	0.41	6.07	l_head	0.14 ± 0.19	[0.11, 0.16]	0.21	2.76
Lengths	0.17 ± 0.46	[0.15, 0.19]	0.46	9.42	Lengths	0.10 ± 0.43	[0.08, 0.12]	0.43	5.80
h_head	0.12 ± 0.18	[0.10, 0.15]	0.18	5.67	h_head	0.11 ± 0.16	[0.08, 0.13]	0.17	4.86
h_1seg	0.10 ± 0.12	[0.08, 0.11]	0.12	4.70	h_1seg	0.08 ± 0.10	[0.07, 0.09]	0.10	3.70
h_2seg	0.08 ± 0.10	[0.06, 0.09]	0.11	3.61	h_2seg	0.07 ± 0.10	[0.05, 0.08]	0.10	3.03
h_3seg	0.07 ± 0.10	[0.06, 0.08]	0.10	3.15	h_3seg	0.07 ± 0.10	[0.06, 0.08]	0.10	2.84
h_4seg	0.08 ± 0.09	[0.07, 0.09]	0.11	3.78	h_4seg	0.06 ± 0.09	[0.04, 0.06]	0.09	2.48
h_5seg	0.09 ± 0.11	[0.08, 0.11]	0.12	4.83	h_5seg	0.07 ± 0.10	[0.06, 0.09]	0.10	3.63
h_6seg	0.04 ± 0.08	[0.03, 0.05]	0.08	2.69	h_6seg	0.05 ± 0.08	[0.04, 0.06]	0.08	2.96
Heights	0.08 ± 0.12	[0.08, 0.09]	0.12	4.06	Heights	0.07 ± 0.11	[0.06, 0.08]	0.11	3.36
w_head	0.13 ± 0.11	[0.13, 0.16]	0.16	8.20	w_head	0.08 ± 0.1	[0.06, 0.09]	0.10	4.87
w_1seg	0.14 ± 0.10	[0.12, 0.15]	0.16	9.44	w_1seg	0.07 ± 0.08	[0.06, 0.07]	0.08	4.50
w_2seg	0.20 ± 0.07	[0.19, 0.21]	0.21	13.97	w_2seg	0.06 ± 0.07	[0.05, 0.06]	0.07	3.82
w_3seg	0.15 ± 0.06	[0.14, 0.17]	0.17	12.32	w_3seg	0.04 ± 0.06	[0.04, 0.05]	0.06	3.42
w_4seg	0.12 ± 0.06	[0.12, 0.14]	0.14	11.48	w_4seg	0.04 ± 0.05	[0.04, 0.05]	0.06	4.02
w_5seg	0.10 ± 0.06	[0.10, 0.12]	0.11	10.03	w_5seg	0.05 ± 0.06	[0.04, 0.05]	0.06	4.75
w_6seg	0.07 ± 0.06	[0.07, 0.09]	0.08	9.20	w_6seg	0.04 ± 0.05	[0.04, 0.05]	0.05	5.58
Widths	0.13 ± 0.09	[0.12, 0.14]	0.15	10.49	Widths	0.05 ± 0.07	[0.05, 0.06]	0.07	4.50
General	0.13 ± 0.28	[0.12, 0.14]	0.28	8.14	General	**0.08 ± 0.25****	[0.07, 0.08]	**0.25**	**4.56**

To validate the observed differences, a paired Wilcoxon signed-rank test was performed on the error distributions of the two methods. The "General" MAE for the SVM regression approach was found to be statistically significant ($p = 0.00039$), confirming its superiority over the non-regression method. This significance is denoted in the table with asterisks (*** $p < 0.001$).

To clarify the aggregation method used in Table 6, the metrics for each "Morphological Variable" (e.g., 'l_head') are calculated by averaging the errors for that specific measurement across all test specimens. The "General" metrics are computed by first pooling all individual measurements (all variables from all specimens) into a single dataset, and then calculating the overall MAE, RMSE, and MAPE from this complete pool. The "General" MAE, therefore, represents the average error expected from any single measurement performed by the Imashrimp system.

Morphological lengths were calculated from keypoints distributed along the shrimp virtual skeleton in both lateral and dorsal point of views (Fig. 6). To determine the optimal point of view for retrieving length variables, the test subset with the **same animals** photographed laterally and dorsally was used to compare which point of view provided the most accurate length measurements. The comparison results are presented in Table 7.

A paired Wilcoxon signed-rank test confirmed that the difference in the general 'Lengths' metric between the two point of views is statistically significant ($p = 0.019$). Based on this result, and to ensure maximum precision for genetic selection, our system is configured to derive length measurements **exclusively** from the Lateral point of view, rather than using Dorsal estimations as a fallback.

Overall we can see that our regression obtains measurements with less than a millimetre of error on average. When assessing the full

Table 7
Comparative error analysis of using Lateral or Dorsal point of view to measure length variables. (Significance test: * $p < 0.05$).

Lateral Point of View Network					Dorsal Point of View Network				
Variable	MAE (cm)	IC 95% (MAE)	RMSE (cm)	MAPE (%)	Variable	MAE (cm)	IC 95% (MAE)	RMSE (cm)	MAPE (%)
l_1seg	0.09 ± 0.11	[0.08, 0.10]	0.11	6.76	l_1seg	0.10 ± 0.11	[0.09, 0.11]	0.11	7.05
l_2seg	0.07 ± 0.09	[0.06, 0.08]	0.09	6.18	l_2seg	0.07 ± 0.10	[0.06, 0.09]	0.10	6.80
l_3seg	0.09 ± 0.12	[0.07, 0.10]	0.12	6.29	l_3seg	0.09 ± 0.13	[0.08, 0.11]	0.13	6.70
l_4seg	0.09 ± 0.11	[0.08, 0.10]	0.11	8.34	l_4seg	0.09 ± 0.11	[0.08, 0.10]	0.11	8.45
l_5seg	0.20 ± 1.10	[0.05, 0.30]	1.11	7.79	l_5seg	0.19 ± 1.09	[0.06, 0.30]	1.09	7.88
l_6seg	0.06 ± 0.08	[0.06, 0.07]	0.08	3.25	l_6seg	0.07 ± 0.09	[0.07, 0.09]	0.09	3.67
l_head	0.14 ± 0.19	[0.11, 0.16]	0.21	2.75	l_head	0.19 ± 0.26	[0.17, 0.23]	0.26	3.93
Lengths	**0.10*** ± **0.43**	[0.08, 0.12]	**0.43**	**5.80**	Lengths	0.12 ± 0.43	[0.01, 0.14]	0.43	6.32

length estimation of the shrimp the error is also quite low, 5 millimetres, but is affected by the detection of the *rostrum*. This in any case is not very relevant as morphological measurements for the full length but ignoring the *rostrum* yields an much lower overall error of 2 millimetre. It has to be underlined that in the are that we ignore the animal has no meat content.

6.4. Ablation study

To quantify the contribution of our system's core components and validate our architectural choices, we conducted a comprehensive ablation study. We established our baseline model (the full Imashrimp system: ViT-Pose Huge, RGB-D input, and both discriminators active) and systematically ablated three key aspects: (1) the dual discrimination modules, (2) the contribution of the depth (D) channel, and (3) the effect of the pose estimation model size. The results of this study are presented in Table 8, reporting 2D pose estimation performance (mAP and PCK). We now describe in detail the results for each of the ablated parameters:

Impact of Discrimination Modules: The first set of experiments (Rows 1–3) demonstrates the critical role of the discriminators. Disabling both the point of view (POV) and rostrum integrity (RI) modules causes the mAP to drop significantly from 96.84% to 91.62%, as the system attempts to process images with the incorrect pose estimation models.

Disabling only the RI, or only the POV, modules also results in a substantial performance degradation (93.53% mAP and 92.62% mAP, respectively). It is worth noting that the ablation of the discriminator module implies leaving the initial human error, that was quantified in Table 2.

Impact of Depth Channel (RGB-D vs. RGB): The resulting mAP dropped from 96.84% to 90.81%. This shows the addition of depth information to be a godd regularizer for keypoint detection in our case.

Impact of Model Size: We compared our ViTPose Huge model against its smaller variants (Large, Base, and Small). The ViTPose Huge model (96.84% mAP) significantly outperforms all smaller architectures, which achieved mAP scores clustering between 89.89% and 90.70%. This result justifies our choice of the larger "Huge" architecture, indicating that the complexity of the shrimp pose estimation task benefits from the increased model capacity.

6.5. End-to-end results

Quantitative results of the complete system can be seen in Table 6, while qualitative results are shown in Fig. 8 where keypoint outputs are combined into a skeleton and visually represented. Furthermore, the size regressor converts pixel-based morphological variables into centimetres.

To demonstrate the qualitative robustness and versatility of the pose estimation system, predictions were performed on a dataset with scenarios not included in training. These include images varying backgrounds and different camera distances. Examples of these results are presented in Fig. 9.

7. Discussion and conclusion

The primary contribution of this work is the introduction of Imashrimp, an artificial intelligence (AI) system comprising multiple integrated modules designed to automate morphological analysis of white shrimp (*Penaeus vannamei*) and minimize human error. To our knowledge, no previous studies have applied body pose estimation techniques to shrimp or performed regression analysis of 23 morphological variables using RGB-D images. Another key feature is its dual discrimination modules, which learn to identify shrimp point of view (lateral or dorsal) and rostrum integrity (good or broken), which, as shown, significantly reduces human annotation error.

The results are promising and could improve genetic selection by automating phenotypic analyses, enabling larger population studies with fewer errors. For the complete test subset of 1245 images, discriminators operating as a two-factor authentication system reduced human error in point of view discrimination from 0.64% to 0% and in rostrum integrity discrimination from 10.44% to 1.04%.

The pose estimation module demonstrated robust performance, which served as a precise foundation for the subsequent 2D-to-3D measurement regression. Our SVM-based conversion module successfully retrieved morphological variables, achieving a general Mean Absolute Error (MAE) of 0.08 ± 0.25 cm.

Obtaining such high precision 3D measurements opens a wide array of possibilities. Imashrimp will greatly increase the number of specimens that are considered for genetic selection, making the genetic selection much more robust and providing significant economic benefits to aquaculture companies that employ such system. Specifically, manual phenotyping of 23 variables on dead specimens typically requires two operators and approximately 9 min per animal, while even a simplified analysis on live animals (4 variables) takes about 2 min. In contrast, Imashrimp requires only a single operator for image acquisition and processes the comprehensive morphometric data in just 32 ms per image, representing a massive reduction in labour costs and time. In terms of operational efficiency, the system achieves an inference speed of approximately 31 Frames Per Second (FPS) (32 ms per image) on a workstation with a dedicated Graphics Processing Unit (GPU) (see Appendix A), demonstrating its feasibility for real-time high-throughput applications in industrial breeding facilities..

8. Limitations and future work

Despite this performance, we acknowledge several limitations that define the scope of this work:

First, the model was trained and validated exclusively in a controlled laboratory setting with uniform backgrounds and lighting. Its robustness to 'in-the-wild' conditions, such as the variable lighting, water reflections, and occlusions found in industrial processing plants, has not yet been quantified experimentally (e.g., via PCK drop analysis) due to the lack of a ground-truth annotated dataset for these unstructured environments. Therefore, the current "in-the-wild" results (Fig. 9) remain qualitative demonstrations of potential transferability.

Table 8
Ablation study evaluating the impact of discriminators Point of View/Rostrum Integrity (POV/RI), input data (RGB-D vs. RGB), and model size (ViTPose variant) on pose estimation accuracy.

Discrimination		Pose Estimation			
POV	RI	Input	Model	mAP 50:95 (%)	PCK@10px (%)
False	False	RGB-D	ViT-Pose Huge	91.62	89.88
True	False	RGB-D	ViT-Pose Huge	93.53	91.45
False	True	RGB-D	ViT-Pose Huge	92.62	90.75
True	True	RGB	ViT-Pose Huge	90.81	82.98
True	True	RGB-D	ViT-Pose Large	90.09	82.83
True	True	RGB-D	ViT-Pose Base	89.89	83.01
True	True	RGB-D	ViT-Pose Small	90.70	83.39
True	True	RGB-D	ViT-Pose Huge	**96.84**	**91.67**

Fig. 8. Test results for the whole proposed system, in which classify information, keypoints and specimen morphological variables in centimetres are shown for each detected shrimp instance. The images show several results of the test dataset depending on the Pose Estimation network by which they have been processed: **(1) Lateral Pose Estimation - 23 keypoints**, **(2) Dorsal Pose Estimation - 23 keypoints**, **(3) Lateral Pose Estimation - 22 keypoints** and **(4) Dorsal Pose Estimation - 22 keypoints**.

Second, the current system is dependent on 4-channel RGB-D data to achieve its reported accuracy. This dependence on specialized cameras limits its immediate applicability in settings equipped only with standard RGB cameras. Future work will focus on bridging this performance gap, potentially through domain adaptation or advanced data augmentation, to create a robust system that relies solely on standard RGB imagery.

Second, the current system is dependent on 4-channel RGB-D data to achieve its reported accuracy. This dependence on specialized cameras limits its immediate applicability in settings equipped only with standard RGB cameras. While utilizing standard RGB cameras would reduce hardware costs, our ablation results indicate that this significantly compromises performance (dropping from 96.84% to 90.81% mAP) while offering a negligible improvement in inference speed (31 ms vs. 32 ms). Consequently, to maintain the precision required for genetic selection, the current minimum viable hardware configuration for field adoption must include a depth-sensing (RGB-D) capability and a CUDA-enabled GPU. Future work will focus on bridging this performance gap, potentially through domain adaptation or advanced data augmentation, to create a robust system that relies solely on standard RGB imagery.

Third, the model was trained and validated exclusively on *Penaeus vannamei*. While generalization to morphologically distinct species would require a complete re-annotation, we hypothesize that the IMASHRIMP framework could be **directly applied** to other commercially relevant species within the same *Litopenaeus* subgenus (e.g., *Litopenaeus stylirostris* or *Litopenaeus setiferus*) due to their similar morphology. Consequently, we anticipate that extending the system to these related species would require either zero or, at most, limited labelling (transfer learning) for validation, rather than a full dataset overhaul.

Finally, as the dataset is proprietary due to commercial confidentiality, external reproducibility is limited, a factor we have aimed to mitigate through detailed methodological descriptions. These limitations also highlight future applications, such as employing Imashrimp in packaging plants where manual measurements are currently infeasible. This would provide companies with richer data for strategic planning.

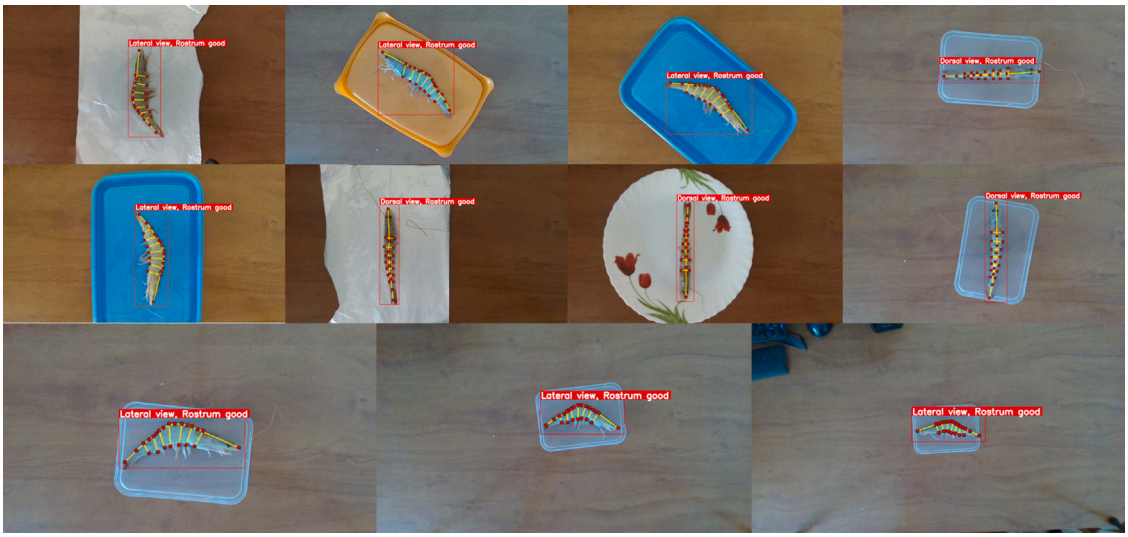

Fig. 9. Evaluation of the proposed system *in-the-wild*, The proposes system is tested in an array of situations not seen during training and classification information and keypoints are shown for each detected shrimp instance. The images show several results depending on the experiment: **First and second row:** Different backgrounds at the same distance and **Third row:** Different distances to the plane 30 cm, 40 cm and 60 cm.

CRediT authorship contribution statement

Remache González Abiam: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chagour Meriem:** Investigation, Data curation. **Bijan Rüth Timon:** Investigation, Data curation. **Trapiella Cañedo Raúl:** Validation, Data curation. **Martínez Soler Marina:** Writing – review & editing, Resources, Investigation, Data curation. **Lorenzo Felipe Álvaro:** Validation, Methodology, Investigation. **Shin Hyun-Suk:** Resources. **Zamorano Serrano María-Jesús:** Resources. **Torres Ricardo:** Funding acquisition. **Castillo Parra Juan-Antonio:** Funding acquisition. **Reyes Abad Eduardo:** Funding acquisition. **Ferrer Ballester Miguel-Ángel:** Funding acquisition. **Afonso López Juan-Manuel:** Supervision, Project administration. **Hernández Tejera Francisco-Mario:** Writing – review & editing, Supervision. **Penate-Sanchez Adrian:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Ethical statement

All shrimp (*Penaeus vannamei*) used in this study were sourced from the ongoing PMG-BIOGEMAR genetic breeding program (Shin et al., 2020). The animals were handled and sacrificed in accordance with the standard commercial aquaculture practices used for human consumption, and no live animals were used in this study. As this research was based on analysing images of animals sampled from this routine commercial process, no additional experimental procedures were performed on live animals.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used the AI-assisted tool **Overleaf Widscribe** to improve the clarity and readability of some sentences, mainly to make the text more suitable for an academic research article. After using this tool, the authors reviewed and edited the content as necessary, and take full responsibility for the content of the publication.

Fundings

This research was conducted as part of a genetic improvement program for whiteleg shrimp (**Penaeus vannamei**), carried out in collaboration between the University of Las Palmas de Gran Canaria (ULPGC) and the Ecuadorian company BIOGEMAR S.A., part of the ALMAR Group. The project is supported by the research grant **C2021/72 "Implementación de un programa de mejora genética para la producción del langostino blanco BIOGEMAR"**, within a long-term agreement focused on the development and application of BLUP-based selection methodologies to enhance shrimp growth performance while maintaining environmental sustainability. The partnership includes technological collaboration in software development and has been active since 2014 as part of a broader knowledge transfer framework between ULPGC and BIOGEMAR.

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.

Appendix A. Implementation, dataset, and hardware details

To ensure methodological transparency and aid reproducibility, this appendix provides detailed information on the dataset, training hyperparameters, and the hardware environment used for all experiments.

A.1. Hardware and inference performance

All training and inference benchmarks were conducted on the hardware configuration detailed in Table A.9. The inference performance of the final, end-to-end Imashrimp system (including discrimination, pose estimation, and regression) is reported in Table A.10.

A.2. Dataset details

We now detail the whole process of how the dataset was created, the tools used, the precise process, and the final characteristics of the constructed dataset.

Table A.9
Hardware Specifications. Abbreviations: CPU: Central Processing Unit; GB: Gigabytes; GPU: Graphics Processing Unit; RAM: Random Access Memory; VRAM: Video Random Access Memory.

Component	Specification
CPU	AMD Ryzen 9 7950X 16-Core Processor
GPU	NVIDIA GeForce RTX 3090
VRAM	24 GB
System RAM	64 GB
Operating System	Windows 10 Enterprise 64-bit

A.2.1. Data acquisition and labelling protocol

As described in Section 5.1, all images were captured in a controlled laboratory setting with constant, uniform lighting and a consistent black background.

The labelling protocol involved several stages to ensure high-quality annotations:

1. **Initial Training**: Annotators were instructed by morphological experts on the 23 keypoints and the use of the CVAT annotation tool.
2. **Manual Annotation**: The initial batch of images was fully annotated manually.
3. **AI-Assisted Annotation**: Once the pose estimation models achieved sufficient accuracy, an auto-labelling pipeline was implemented. Annotators then focused on correcting minor inaccuracies, significantly speeding up the process.
4. **Expert Review**: Crucially, every annotation (both manual and corrected) was subsequently reviewed and validated by a morphological expert to ensure correctness and consistency. This two-step process (annotation + expert review) served to maximize inter-annotator reliability.

A.2.2. Dataset split and class balance

The full dataset of 12,367 images was split into training (80%), validation (10%), and test (10%) sets. This split was performed at the **specimen level**, ensuring that no single shrimp appears in more than one set, which validates the model's ability to generalize to unseen individuals.

The specific datasets for the pose estimation models were derived from this split as follows:

- 23-Keypoint Models (e.g., lateral-23): To train the models for the complete skeleton, only images where the rostrum was intact ("rostrum good") were selected. Images corresponding to "rostrum broken" were excluded from this dataset, as they physically lack keypoint 1.
- 22-Keypoint Models (e.g., lateral-22): To train the models that operate without the rostrum tip, the entire dataset (12,367 images) was used, including both "rostrum good" and "rostrum broken" images. For this dataset, the ground truth was adapted by systematically removing keypoint 1 from all annotations.

The class balance for the discriminator modules and the resulting image counts for each pose estimation model are detailed in Tables A.11 and A.12, respectively.

A.3. Training hyperparameters and data augmentation

All models were trained using the parameters specified in Table A.13. The pose estimation models were initialized from ViTPose Huge weights pre-trained on MS-COCO/MPII, with the input layer modified to accept 4 channels (RGB-D).

Data Augmentation (Pose Estimation) To ensure robustness, the following augmentation pipeline was applied sequentially during the training of all pose estimation models:

Table A.10
System Performance and Resource Usage. Abbreviations: FPS: Frames Per Second, ms: milliseconds.

Metric	Training	Test
Inference Time per Image	140 ms	32 ms
Frames Per Second	7 FPS	**32 FPS**
VRAM Usage	22 GB	8 GB
GPU Core Load	82%	35%
CPU Load	20%	20%

Table A.11
Class balance for discriminator modules.

Module	Class	Train	Validation	Test	Total
Point of View	Lateral	6816	821	831	8468
	Dorsal	3079	406	414	3899
Rostrum Integrity	Good	8700	1047	1017	10764
	Broken	1195	180	228	1603
Total Images		**9895**	**1227**	**1245**	**12367**

Table A.12
Image counts for pose estimation models.

Model	Train	Validation	Test	Total
Lateral-23	5997	701	679	7377
Lateral-22	6816	821	831	8468
Dorsal-23	2703	346	338	3387
Dorsal-22	3079	406	414	3899

- **Image Loading:** Load the 4-channel RGB-D image.
- **Geometric Augmentation I (Flip):** Apply a top-down random flip with a 50% probability.
- **Geometric Augmentation II (Half-Body):** Apply a half-body transform, focusing on a subset of 8 keypoints with a 30% probability.
- **Geometric Augmentation III (Scale/Rotation):** Apply random scaling (up to 0.5 scale factor) and random rotation (up to 40 degrees).
- **Affine Transform:** Apply the final affine transformation based on the geometric augmentations.
- **Tensor Conversion:** Convert the augmented image to a tensor.
- **Normalization:** Normalize the tensor using pre-calculated `mean` and `std` values.
- **Target Generation:** Generate the target heatmaps from keypoint coordinates (using a sigma of 2).
- **Data Collection:** Collect the final keys required for model training.

Appendix B. Detailed 2D error and confidence score analysis

This section provides a detailed, stratified analysis of the 2D pose estimation performance. The pixel predictions from the test set were compared to the annotated ground truth to provide a granular 2D error analysis. These errors are presented in Tables B.14 and B.15, showing the performance for each keypoint across all four specialized sub-models (Lateral-23, Lateral-22, Dorsal-23, and Dorsal-22). This table includes standard 2D error metrics (EPE (px), RMSE (px), and MAPE (%)) and a 'Score' column, which reports the average heatmap confidence (i.e., heatmap variance) for each keypoint as a measure of the model's prediction certainty.

Data availability

The authors do not have permission to share data.

Table A.13
Training hyperparameters. Abbreviations: MSE: Mean Square Error; BCE: Binary Cross-Entropy; BS: Batch Size; LR: Learning Rate; Funct.: Function; ViT-Pose-H: Vit-Pose-Huge.

Module	Architecture	Optimizer	Loss Funct.	LR	BS	Epochs	Checkpoint Criterion
Point of View Disc.	ResNet-50	Adam	BCELoss	0.00200	256	5	Best Validation Error
Rostrum Disc.	ResNet-50	Adam	BCELoss	0.00050	256	5	Best Validation Error
Pose Est. (Dorsal-22)	RGBD ViTPose-H	Adam	JointsMSELoss	0.00010	16	210	Best Validation Loss
Pose Est. (Dorsal-23)	RGBD ViTPose-H	Adam	JointsMSELoss	0.00070	16	210	Best Validation Loss
Pose Est. (Lateral-22)	RGBD ViTPose-H	Adam	JointsMSELoss	0.00007	16	210	Best Validation Loss
Pose Est. (Lateral-23)	RGBD ViTPose-H	Adam	JointsMSELoss	0.00007	16	210	Best Validation Loss

Table B.14
Comparative 2D error analysis for Dorsal sub-models (Dorsal-22 and Dorsal-23).

Point	Dorsal-22 sub-model				Dorsal-23 sub-model			
	EPE (px)	RMSE (px)	MAPE (%)	Score	EPE (px)	RMSE (px)	MAPE (%)	Score
1	–	–	–	–	12.44 ± 15.71	14.17	1.17	0.71
2	3.18 ± 2.23	2.74	0.30	0.95	3.77 ± 2.10	3.05	0.35	0.93
3	3.46 ± 2.02	2.84	0.32	0.94	3.41 ± 1.91	2.76	0.31	0.94
4	4.05 ± 2.62	3.41	0.36	0.93	4.04 ± 2.76	3.46	0.36	0.92
5	5.26 ± 3.57	4.50	0.45	0.91	5.04 ± 3.22	4.23	0.45	0.93
6	2.98 ± 1.62	2.40	0.27	0.93	3.36 ± 1.97	2.75	0.31	0.94
7	3.07 ± 1.69	2.48	0.30	0.95	3.21 ± 1.80	2.60	0.31	0.93
8	2.89 ± 1.63	2.35	0.28	0.98	2.97 ± 1.70	2.42	0.30	0.93
9	7.15 ± 4.13	5.84	0.65	0.80	8.21 ± 3.68	6.36	0.85	0.82
10	8.61 ± 9.17	8.89	0.72	0.81	9.79 ± 5.21	7.84	0.84	0.75
11	8.74 ± 9.56	9.16	0.76	0.83	9.58 ± 5.55	7.83	0.83	0.76
12	4.02 ± 2.88	3.50	0.36	0.96	4.08 ± 2.64	3.44	0.35	0.95
13	3.85 ± 2.73	3.34	0.34	0.95	4.14 ± 2.71	3.50	0.37	0.96
14	3.42 ± 2.51	3.00	0.30	0.96	3.91 ± 2.29	3.20	0.34	0.96
15	3.30 ± 2.14	2.78	0.30	0.95	3.75 ± 2.32	3.12	0.33	0.96
16	3.78 ± 2.62	3.25	0.32	0.96	4.20 ± 4.05	4.13	0.37	0.95
17	3.78 ± 2.54	3.22	0.32	0.95	4.11 ± 2.76	3.50	0.36	0.94
18	3.76 ± 2.58	3.22	0.32	0.94	4.10 ± 2.70	3.47	0.37	0.96
19	3.67 ± 2.35	3.08	0.31	0.94	3.91 ± 2.67	3.35	0.34	0.95
20	3.09 ± 2.08	2.64	0.27	0.96	3.41 ± 2.99	3.21	0.32	0.96
21	3.11 ± 2.01	2.62	0.28	0.96	3.34 ± 2.89	3.13	0.31	0.96
22	3.96 ± 2.68	3.38	0.38	0.94	4.51 ± 3.06	3.85	0.43	0.93
23	3.98 ± 2.74	3.42	0.37	0.95	4.44 ± 3.21	3.87	0.42	0.93
General 22KP	4.23 ± 4.08	4.15	0.38	0.93	4.52 ± 3.15	3.90	0.42	0.89
General 23KP	–	–	–	–	4.87 ± 4.78	4.82	0.45	0.92

Table B.15
Comparative 2D error analysis for Lateral sub-models (Lateral-22 and Lateral-23).

Point	Lateral-22 sub-model				Lateral-23 sub-model			
	EPE (px)	RMSE (px)	MAPE (%)	Score	EPE (px)	RMSE (px)	MAPE (%)	Score
1	–	–	–	–	11.66 ± 35.24	26.25	0.88	0.74
2	3.01 ± 1.84	2.49	0.28	0.94	2.74 ± 1.53	2.22	0.25	0.96
3	3.66 ± 2.38	3.09	0.32	0.93	2.99 ± 1.90	2.51	0.27	0.94
4	5.12 ± 3.88	4.54	0.45	0.91	3.88 ± 3.10	3.52	0.35	0.93
5	4.80 ± 3.31	4.12	0.43	0.89	4.29 ± 3.11	3.75	0.40	0.93
6	3.27 ± 1.79	2.64	0.31	0.93	3.17 ± 2.11	2.69	0.30	0.94
7	3.19 ± 2.09	2.70	0.32	0.94	2.79 ± 1.75	2.33	0.27	0.95
8	3.40 ± 2.03	2.80	0.35	0.94	3.15 ± 2.06	2.66	0.32	0.95
9	7.62 ± 3.25	5.86	0.78	0.74	7.53 ± 13.54	10.96	0.65	0.80
10	8.97 ± 4.55	7.12	0.76	0.77	9.55 ± 6.03	7.99	0.82	0.78
11	11.10 ± 8.83	10.03	1.02	0.80	6.00 ± 6.58	6.30	0.56	0.87
12	3.69 ± 2.53	3.17	0.33	0.92	3.09 ± 1.89	2.56	0.27	0.94
13	7.76 ± 6.95	7.37	0.70	0.83	4.97 ± 4.58	4.78	0.44	0.92
14	3.97 ± 2.66	3.38	0.36	0.92	3.37 ± 4.52	3.98	0.30	0.94
15	4.78 ± 3.00	3.99	0.43	0.90	3.93 ± 4.70	4.34	0.35	0.93
16	4.59 ± 3.28	3.99	0.40	0.91	3.93 ± 6.13	5.15	0.37	0.93
17	4.53 ± 3.15	3.91	0.41	0.89	3.93 ± 5.45	4.75	0.37	0.93
18	3.90 ± 2.80	3.40	0.37	0.92	3.59 ± 5.49	4.64	0.35	0.94
19	4.77 ± 3.48	4.18	0.44	0.89	4.09 ± 5.10	4.62	0.39	0.93
20	3.35 ± 2.08	2.79	0.33	0.94	2.87 ± 1.71	2.36	0.28	0.94
21	5.55 ± 3.99	4.83	0.52	0.88	4.40 ± 5.08	4.75	0.42	0.93
22	3.41 ± 2.14	2.85	0.36	0.94	3.34 ± 2.14	2.81	0.34	0.96
23	3.42 ± 1.91	2.77	0.33	0.94	3.38 ± 2.10	2.82	0.32	0.95
General 22KP	4.90 ± 4.21	4.57	0.46	0.89	4.14 ± 3.65	3.90	0.38	0.90
General 23KP	–	–	–	–	4.46 ± 8.31	6.67	0.40	0.92

References

Ana, N., Lee, I., Dulce, S., Henríquez, P., Rodríguez, F.A., Morales, A., Soula, M., Rodrigo, B., Negrín-Báez, D., Zamorano, M., Ha, M., 2016. Imafish_ml: A fully-automated image analysis software for assessing fish morphometric traits on gilthead seabream (sparus aurata l.), meagre (argyrosomus regius) and red porgy (pagrus pagrus). Comput. Electron. Agric. 121, 66–73. http://dx.doi.org/10.1016/j.compag.2015.11.015.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 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, IEEE, pp. 3686–3693.

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J., 2019. Yolact: Real-time instance segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 9156–9165.

Çakmak, M., 2025. A new lightweight hybrid model for pistachio classification using transformers and EfficientNet. IEEE Access 13, 85857–85872.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y., 2021. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. 43 (1), 172–186.

Chao Zhou, L.S., Yang, Guowei, 2021. Counting, locating, and sizing of shrimp larvae based on density map regression. Aquac. Fish. 6, 371–380.

Chirdchoo, N., Mukviboonchai, S., Cheunta, W., 2024. A deep learning model for estimating body weight of live pacific white shrimp in a clay pond shrimp aquaculture. Intell. Syst. Appl. 24, 200434.

Climent-Perez, P., Galán-Cuenca, A., Garcia-d'Urso, N.E., Saval-Calvo, M., Azorin-Lopez, J., Fuster-Guillo, A., 2024. Simultaneous, vision-based fish instance segmentation, species classification and size regression. PeerJ Comput. Sci. 10.

Dong, Q., Li, Z., Yu, X., 2023. Detection-regression based framework for fish keypoints detection. IMTS 1–5.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR.

Food and Agriculture Organization of the United Nations (FAO), 2024. The State of World Fisheries and Aquaculture 2024: Blue Transformation in Action. FAO, Rome, Italy, http://dx.doi.org/10.4060/cc9420en.

Garcia, R., Prados, R., Quintana, J., Tempelaar, A., Gracias, N., Rosen, S., Vågstøl, H., Løvall, K., 2019. Automatic segmentation of fish using deep learning with application to fish size measurement. ICES J. Mar. Sci. 77 (4), 1354–1366.

García-Santamaría, R., Pérez-Rodríguez, E., Pascual, D., García, D., 2022. Automatic, operational, high-resolution monitoring of fish length and catch composition from fish auctions using computer vision. Fish. Res. 251, 106226.

Hadiyanto, H., Widodo, C., 2022. Shrimp body weight estimation in aquaculture ponds using morphometric features based on underwater image analysis and machine learning approach. Rev. D'Intell. Artif. 36, 905–912.

Harbitz, A., 2007. Estimation of shrimp (pandalus borealis) carapace length by image analysis. Ices J. Mar. Sci. - ICES J MAR SCI 64, 939–944.

He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2017. Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision. ICCV, pp. 2980–2988.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Huang, K., Li, Y., Suo, F., Xiang, J., 2020. Stereo vison and mask-rcnn segmentation based 3d points cloud matching for fish dimension measurement. In: 2020 39th Chinese Control Conference. CCC, pp. 6345–6350.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2014. Human3.6 m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. 36 (7), 1325–1339.

Jocher, Glenn, Qiu, Jing, 2025. Ultralytics YOLO26 (version 26.0.0). https://github.com/ultralytics/ultralytics.

Kartbak, S.B.A., Özel, M.B., Kocakaya, D.N.C., Çakmak, M., Sinanoğlu, E.A., 2025. Classification of intraoral photographs with deep learning algorithms trained according to cephalometric measurements. Diagnostics 15, 1059.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P., 2015. Microsoft coco: Common objects in context. arXiv:1405.0312.

Liu, Y., Zhang, Y., Liu, S., Yang, X., Xu, Z., Jiang, S., Xu, D., Yuille, A.L., Roth, H.R., 2022. Semi-supervised medical image segmentation using cross-model pseudo-supervision. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Springer International Publishing, pp. 137–147.

Marshall, J., Klibaite, U., Gellis, A., Aldarondo, D., Olveczky, B., Dunn, T.W., 2021. The pair-r24m dataset for multi-animal 3d pose estimation. In: Vanschoren, J., Yeung, S. (Eds.), NeurIPS, In: NeurIPS Track on Datasets and Benchmarks, vol. 1, pp. 1–5.

Martínez Soler, M., Shin, H.S., Lorenzo-Felipe, Álvaro, Zamorano Serrano, M.J., Ginés Ruiz, R., Pachón Mesa, L.C., González, D., Fernández Martín, J., Ramírez Artiles, J.S., Peñate Sánchez, A., Lorenzo Navarro, J., Torres, R., Reyes Abad, E., Afonso López, J.M., Lince, J.A., 2024. Genetic parameters of meat quality, external morphology, and growth traits in Pacific white shrimp (penaeus vannamei) from an Ecuadorian population. Aquaculture 593, 741228. http://dx.doi.org/10.1016/j.aquaculture.2024.741228, https://www.sciencedirect.com/science/article/pii/S0044848624006896.

Moreno-Noguer, F., 2016. 3D human pose estimation from a single image via distance matrix regression. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1561–1570.

Navarro, A., Lee-Montero, I., Santana, D., Henríquez, P., Ferrer, M.A., Morales, A., Soula, M., Badilla, R., Negrín-Báez, D., Zamorano, M.J., Afonso, J.M., 2016. Imafish_ml: A fully-automated image analysis software for assessing fish morphometric traits on gilthead seabream (sparus aurata l.), meagre (argyrosomus regius) and red porgy (pagrus pagrus). Comput. Electron. Agric. 121, 66–73.

Redmon, J., Divvala, S., Girshick, R.B., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 779–788.

Rong, D., Gang, F., 2024. Coordinate-corrected and graph-convolution-based hand pose estimation method. Sensors 24 (22).

Rueegg, N., Zuffi, S., Schindler, K., Black, M.J., 2022. Barc: Learning to regress 3d dog shape from images by exploiting breed information. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3866–3874.

Shin, H., Martínez-Soler, M., Lorenzo-Felipe, A., Montachana-Chimborazo, M., Yugcha, E., Tomalá, M., Mujica-Rodríguez, K., Mero-Panta, E., Otaiza-Mejillón, J., Franco-Chiquito, N., Fernández Martín, J., Zamorano, M., Intriago-Díaz, W., Afonso, J., 2020. Estimas de parámetros genéticos de la calidad morfológica y del crecimiento de Penaeus vannamei en condiciones industriales de cultivo. Aquacultura 58–62.

Shin, H.S., Montachana Chimborazo, M.E., Escobar Rivas, J.M., Lorenzo-Felipe, Álvaro, Martínez Soler, M., Zamorano Serrano, M.J., Fernández Martín, J., Ramírez Artiles, J.S., Peñate Sánchez, A., Lorenzo Navarro, J., Intriago Díaz, W., Torres, R., Reyes Abad, E., Afonso López, J.M., 2023. Genetic parameters for growth and morphological traits of the Pacific white shrimp penaeus vannamei from a selective breeding programme in the industrial sector of Ecuador. Aquac. Rep. 31, 101649. http://dx.doi.org/10.1016/j.aqrep.2023.101649, https://www.sciencedirect.com/science/article/pii/S2352513423001886.

Sigal, L., Balan, A., Black, M., 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. 87, 4–27.

Tome, D., Russell, C., Agapito, L., 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1–5.

Tonachella, N., Martini, A., Martinoli, M., Pulcini, D., Romano, A., Capoccioni, F., 2022. An affordable and easy-to-use tool for automatic fish length and weight estimation in mariculture. Sci. Rep. 12 (1), 1–11.

Vapnik, V., Golowich, S.E., Smola, A., 1996. Support vector method for function approximation, regression estimation and signal processing. In: Proceedings of the 9th International Conference on Neural Information Processing Systems. NIPS'96, MIT Press, Cambridge, MA, USA, pp. 281–287.

Voskakis, D., Makris, A., Papandroulakis, N., 2021. Deep learning based fish length estimation. an application for the Mediterranean aquaculture. In: OCEANS 2021: San Diego – Porto. IEEE, pp. 1–5.

Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1–5.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019a. Detectron2. Available at: https://github.com/facebookresearch/detectron2.

Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., Wang, Y., Wang, Y., 2019b. Large-scale datasets for going deeper in image understanding. In: 2019 IEEE International Conference on Multimedia and Expo. ICME, pp. 1480–1485.

Xiao, B., Wu, H., Wei, Y., 2018. Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 1–5.

Xu, J., Zhang, Y., Peng, J., Ma, W., Jesslen, A., Ji, P., Hu, Q., Zhang, J., Liu, Q., Wang, J., Ji, W., Wang, C., Yuan, X., Kaushik, P., Zhang, G., Liu, J., Xie, Y., Cui, Y., Yuille, A., Kortylewski, A., 2023. Animal3d: A comprehensive dataset of 3d animal pose and shape. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 9099–9109.

Xu, Y., Zhang, J., Zhang, Q., Tao, D., 2022. Vitpose: Simple vision transformer baselines for human pose estimation. arXiv:2204.12484.

Yao, C.-H., Hung, W.-C., Li, Y., Rubinstein, M., Yang, M.-H., Jampani, V., 2022. Lassie: Learning articulated shape from sparse image ensemble via 3d part discovery. In: NeurIPS. pp. 1–5.

Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J., 2021. Lite-hrnet: A lightweight high-resolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10440–10450.

Zhou, H., Kim, S.-H., Kim, S.-C., Kim, C.-W., Kang, S.-W., 2023. Size estimation for shrimp using deep learning method. Smart Media J. 12 (3), 112–119.