Deep Learning-Based Methods for Automated Estimation of Insect Length, Volume, and Biomass

SHIRALI, Hossein[1*], ASCENZI, Aleida[2,3], WÜHRL, Lorenz[1], BEYER, Nils[1], DI LORENZO, Noemi[3,4], VACCARELLA, Emanuele[5], KLUG, Nathalie[1], MEIER, Rudolf[6,7], CERRETTI, Pierfilippo[2,3], PYLATIUK, Christian[1]

[1]Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), 76149 Karlsruhe, Germany.

[2] Department of Biology and Biotechnologies "Charles Darwin", Sapienza University of Rome, 00185 Rome, Italy.

[3] Museum of Zoology (MZUR), Sapienza University of Rome, 00162 Rome, Italy.

[4] Department of Life Sciences, University of Siena, 53100 Siena, Italy.

[5] Department of Environmental Biology, Sapienza University of Rome, 00185 Rome, Italy.

[6] Center for Integrative Biodiversity Discovery, Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Invalidenstraße 43, 10115 Berlin, Germany.

[7] Institute of Biology, Humboldt University, 10115 Berlin, Germany.

*Corresponding author: Hossein Shirali (*hossein.shirali@kit.edu*)

Abstract

Accurate information on insect biomass and size is fundamental for studying insect behavior, ecology, and decline. However, most methods are labor-intensive, often invasive, and impede large-scale studies. Here, we introduce two novel, non-invasive, deep learning-based methods to automatically measure key insect traits from images. First, we introduce a general approach using Oriented Bounding Boxes (OBB) designed for broad applicability across diverse insect taxa. By adjusting for specimen orientation, this method measures length accurately with a mean absolute error (MAE) of 0.211 mm compared to conventional photomicroscope measurements and provides initial biomass estimates using established length-weight relationships. Second, we tested whether a specialized segmentation model requiring taxon-specific training improves biomass predictions. We show that a model for Tachinidae (Diptera, Calyptratae) can effectively delineate key body parts (head, thorax, abdomen) and provide accurate curvilinear body length, volume, and biomass estimates, correlating strongly with both wet (R = 0.937) and dry (R = 0.907) weight. Validation experiments demonstrate that our methods are accurate and offer advantages over traditional techniques by reducing handling and improving scalability. This is important because developing similar segmentation models for just the 20 most dominant flying insect families collected in bulk samples like Malaise traps could cover around 50% of specimens, and the modular design of our segmentation pipeline allows for seamless updates to include additional families. Together, these two new approaches for morphometric analysis provide an efficient, scalable framework for automated insect data acquisition, with broad implications for diverse ecological and evolutionary studies, ranging from functional trait ecology and population-level assessments to large-scale biodiversity monitoring and conservation initiatives.

Keywords: Automated morphometrics; Body Length Measurement; Body Volume Estimation; Deep Learning; Ecological Monitoring; Insect biomass; Oriented Bounding Box (OBB); Segmentation-based Method

1. Introduction

Accurate measurement of insect traits, such as length, volume, and biomass, is essential for understanding ecosystem health, species distribution, and ecological dynamics. These functional traits provide crucial insights into a wide array of biological processes, from food webs and energy flow (Cardinale et al., 2012; Loreau et al., 2001) to how insect populations respond to environmental changes like climate shifts (Schowalter, 2016; Tilman et al., 2006) and widespread habitat alteration (Lister & Garcia, 2018).

The applications of morphometric data are extensive and play a key role across multiple research fields. In functional trait ecology, these data help link biomass to dispersal ability and metabolic rate (Bosch & Vicens, 2002; Jahant-Miller et al., 2022). In pollination biology, body size measurements are used to assess the relationship between insect size and pollen transport efficiency (Büyükyilmaz & Tseng, 2022). In pest management, morphometrics contribute to inform on pest control (Maurey et al., 2025), and in behavioral ecology, precise individual-level measurements are essential for studying life history evolution (LaBarbera, 1989) and parasitoid-host dynamics (Poulin & George-Nascimento, 2007; Hechinger, 2013). Furthermore, in biodiversity research, tracking shifts in body size over time and space can be relevant to understanding population dynamics. Indeed, the well-documented decline in insect biomass across ecosystems (Habel et al., 2019; Hallmann et al., 2017; Sánchez-Bayo & Wyckhuys, 2019;

van Klink et al., 2020) further highlights the urgency of developing accurate and efficient morphometric tools to monitor these changes and inform conservation strategies.

However, obtaining comprehensive morphometric data, particularly for the large sample volumes generated by standardized collection methods like Malaise traps, remains a significant challenge (Bazzo et al., 2023; Pellegrino et al., 2022). These traps can yield thousands of specimens from hundreds of species weekly (Karlsson et al., 2020; Srivathsan et al., 2019), often dominated by hyperdiverse groups whose taxonomy is poorly understood (so-called "dark taxa") (Hartop et al., 2022; Page, 2016; Srivathsan et al., 2023). Traditional morphometric methods, typically involving direct weighing after drying, are precise but inherently invasive, labor-intensive, and time-consuming (Rodríguez-Lozano et al., 2021). Furthermore, manual measurements can suffer from inter-observer variability (Rogers et al., 1976; Sample et al., 1993). These limitations render traditional approaches impractical for large-scale studies and preclude subsequent analyses like DNA barcoding, which are increasingly vital for biodiversity assessment (Hebert et al., 2003; Wang et al., 2018).

To address these challenges, image-based automated approaches are being developed. Notably, Schneider et al., 2022 introduced a deep learning method for biomass estimation by segmenting individual arthropods within bulk samples from images. While scalable, it requires calibration data for varied arthropod groups and can face accuracy limitations with morphologically similar species. Other systems like BIODISCOVER (Ärje et al., 2020) estimate biomass using image area as a proxy, enabling high throughput but facing challenges with smaller species due to inconsistent area-to-weight relationships. The DiversityScanner (Wührl et al., 2022) integrates robotics with machine learning for

biomass estimation and sorting by detecting contours and segmenting body parts from 2D images to approximate 3D volumes, though it can require manual adjustments for complex morphologies. While these studies demonstrate the feasibility of image-based morphometrics, challenges persist in achieving high accuracy and adaptability across the diverse morphologies of insect groups, especially when non-invasive, specimen-level data is required.

The field is rapidly moving towards integrated pipelines that combine advancements in robotics, imaging, high-throughput sequencing (HTS), and machine learning to tackle large-scale insect analysis (Wägele et al., 2022). This includes automated specimen handling (Wührl et al., 2022), size-sorting (Ascenzi et al., 2025), high-resolution imaging (Hereld et al., 2017; Klug et al., 2024), and AI-driven classification (Knyshov et al., 2021; Shirali et al., 2024; Caruso et al., 2025). The "reverse workflow" in molecular studies, where specimens are barcoded first and then morphologically validated (Hartop et al., 2024; Srivathsan et al., 2021; Yeo et al., 2021), further emphasizes the need for efficient, non-destructive morphometric tools that can be applied to individuals post-identification or alongside molecular processing. Within this vision of a "next wave" of specimen-based biodiversity assessment, our work aims to provide a crucial missing component: robust, automated, and non-invasive methods for acquiring detailed morphometric data from individual insect images.

Here, we address the need for improved morphometric tools by developing and validating two complementary, non-invasive, deep learning-based methods for estimating insect length, volume, and subsequently biomass from 2D images. Both insect length and volume were chosen as primary targets because their correlation with dry mass is well-established

(Ganihar, 1997; Penell et al., 2018; Sample et al., 1993; Smock, 1980), allowing biomass estimation through non-invasive, automated measurements. Our first method employs Oriented Bounding Boxes (OBB) to provide a general, rapid technique for length estimation across a wide range of insect taxa, adaptable to diverse morphologies and suitable for quick processing of large, morphologically varied samples. Our second, more specialized method utilizes instance segmentation, demonstrated with Tachinidae (Diptera, Calyptratae), to accurately delineate key body parts (head, thorax, abdomen). This allows for detailed curvilinear length and volumetric calculations, providing precise data valuable for in-depth studies of specific taxa. Although developing specialized morphometric models requires taxon-specific training data, this effort can be optimized by leveraging the typical rank-abundance distribution observed in insect communities, where a few dominant families and species account for the majority of individuals in a sample (Srivathsan et al., 2023). Focusing on these common taxa allows models to generate detailed morphometric data for a substantial portion of bulk samples, making this a strategic and efficient approach for large-scale analyses of diversity and biomass.

By developing these novel methodologies, this research contributes to the growing toolkit for automated insect analysis. These approaches enhance the precision and adaptability needed for addressing diverse ecological questions at scale, supporting applications from large-scale biodiversity monitoring to detailed studies of individual species' responses to environmental change.

2. Materials and Methods

2.1 Data Collection and Preparation

2.1.1 Data Acquisition

For the OBB method, 815 images from multiple insect families representing Diptera, Hymenoptera, and Coleoptera were acquired. Images were captured using the low-cost, high-resolution DIY microscope, Entomoscope (Wührl et al., 2024), and stacked using Helicon Focus (*Helicon Soft Ltd*, 2025). Each specimen was preserved in ethanol to maintain its structural integrity during imaging, ensuring consistent quality across the dataset. For the segmentation method, we used bristle flies (Diptera: Tachinidae) as a case study model taxon. We selected 1,320 images from several representative tachinid species for the final training dataset. This family was chosen due to its abundance in samples and distinct morphological features. For both methods, the dataset included images captured from different views and orientations, providing diversity.

Specimen details and image lists are provided in Tables S1 and S2 (Supporting Information). All images and corresponding annotations used in this study are publicly available via (Shirali & Ascenzi, 2025). The datasets were split 70:15:15 by specimen ID to prevent data leakage.

2.1.2 Image Annotation

2.1.2.1 OBB Annotation

For the OBB method, images were labeled using LabelStudio (*HumanSignal/Label-Studio*, 2019/2024). This involved drawing one bounding box around the specimens' main body parts—head, thorax, and abdomen—while accurately capturing their orientation. Non-essential body parts, such as wings, legs, and large antennae, were deliberately excluded to focus on elements critical for length measurement. In cases where

the insect was too large or close to the image borders, making it challenging to label using an oriented bounding box, a 10% padding was added to the borders of all images while maintaining the same background color. This adjustment ensured that the bounding boxes accurately aligned with the insects' natural poses, thereby improving the reliability of length estimation.

2.1.2.2 Segmentation Annotation

We adopted a two-stage annotation strategy to accurately label insect body parts (head, thorax, abdomen) for robust segmentation supporting length and volume estimation. Initially, 820 images were annotated using Grounding DINO (Liu et al., 2023) for bounding boxes and SAM (Kirillov et al., 2023) for segmentation, followed by manual refinement in LabelMe (Wada, 2024). This stage focused only on clearly visible regions, excluding parts obscured by wings or legs, which limited the model's performance.

Due to limitations in capturing complex insect morphology, we improved our approach by annotating 500 additional images using pseudo-labeling from a segmentation model trained on the initial dataset, followed by detailed manual refinement. This refinement explicitly included inferred outlines of body parts partially obscured by legs or wings. This refined method provided a more comprehensive representation of insect anatomy, significantly improving segmentation accuracy and robustness. See Figure 1 for annotation examples.

Figure 1. Annotation strategies for the OBB method (left), targeting the overall specimen outline, and the segmentation method (right), delineating specific body parts.

2.2 Oriented Bounding Box (OBB) Based Length Estimation

2.2.1 OBB Model

To detect oriented bounding boxes aligned with specimen orientation for accurate length measurement, we adapted and evaluated the YOLOv8 OBB architecture (Jocher et al., 2022/2023). This architecture is an advanced version of the YOLO architecture, specifically designed to capture objects' orientations in addition to their positions. Starting with models pre-trained on the general DOTA (Dataset for Object Detection in Aerial Images) dataset, we performed transfer learning by fine-tuning them on our custom-annotated insect dataset to adapt the architecture for our specific morphometrics application. YOLOv8 OBB offers five variants that differ in architectural size and complexity. Generally, larger variants possess greater capacity for achieving higher accuracy but demand more computational resources. Therefore, to identify the optimal balance of performance and efficiency for our application, we systematically trained and evaluated all five variants under identical conditions. The training conditions included a

batch size of 16, AdamW optimizer (learning rate = 0.01), and data augmentations (random rotations, flipping, scaling, mosaic) for improved robustness. The models' performance in bounding box detection and orientation accuracy was assessed using standard metrics: mAP@0.5 (mean Average Precision at 50% IoU overlap) and the stricter mAP@0.5-0.95 (average mAP across IoU thresholds from 0.5 to 0.95). Our evaluation across the variants revealed that the medium version (YOLOv8m-obb) demonstrated the best trade-off between accuracy and computational requirements for this task. Consequently, this variant was selected as the final choice for subsequent length estimation experiments. Figure 2 illustrates the training performance for this selected YOLOv8m-obb model. The evaluation metrics showed steady precision improvement and declining box loss curves during training, indicating effective learning and generalization.

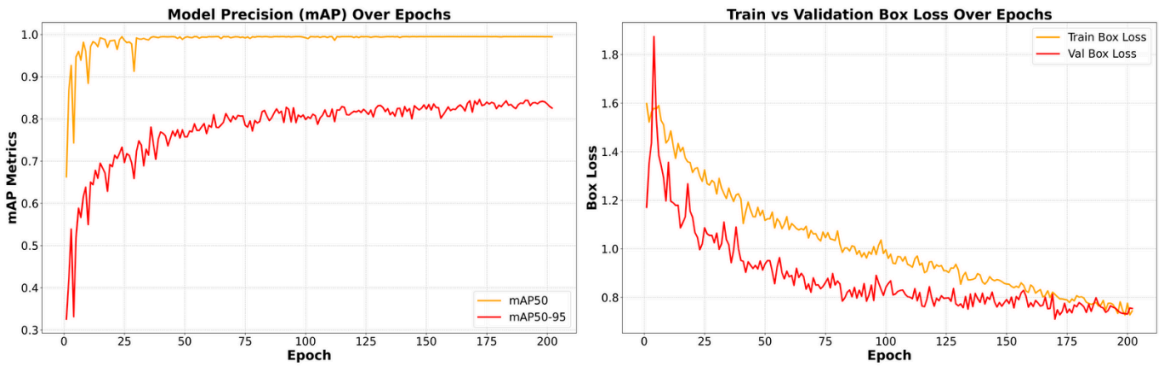

Figure 2. Training performance of the YOLOv8m-obb model, showing convergence of mAP (left) and decreasing box loss (right), indicating effective learning for OBB detection.

2.2.2 Length Measurement from OBB

The bounding box detected by YOLOv8 OBB was used to calculate the specimen's length. The length of the specimen in pixels was determined by identifying the longest side of the bounding box. This pixel length was then converted into millimeters using the same calibration factor derived from the reference object in the segmentation-based method (see

Section 2.4.1). This consistency in unit conversion ensured comparability between the two methodologies.

While the primary focus was on length measurement, the width of the bounding box could also be measured. However, due to variability in specimen poses and potential occlusions, the width measurements are less reliable and were not emphasized in this study. To illustrate the process, Figure 3 showcases an example of a detected bounding box with arrows indicating the measured length and width.

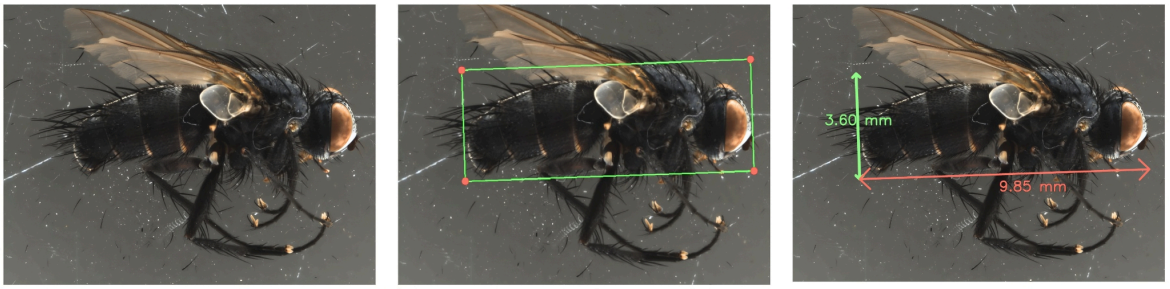

Figure 3. Example output of the OBB method: input image (left), predicted oriented bounding box (center), and derived length (red) and width (green) measurements (right).

2.3 Segmentation-Based Volume and Length Estimation

2.3.1 Body Part Segmentation

To identify the optimal segmentation architecture, Mask R-CNN (He et al., 2018) (ResNet-50-FPN backbone) and YOLOv8-seg (Jocher et al., 2022/2023) (nano variant) were trained and evaluated under identical conditions using the initial 820 labeled images. Both models, pre-trained on the COCO dataset, were subjected to the same augmentations, with performance evaluated using mAP@0.5 and mAP@0.5-0.95 metrics. YOLOv8n-seg demonstrated higher performance, achieving segmentation and bounding box mAP@0.5

scores of 0.944 and 0.964, respectively, versus Mask R-CNN's 0.922 and 0.926. This trend persisted for mAP@0.5–0.95, confirming YOLOv8-seg as the preferred model for subsequent analysis.

YOLOv8-seg also offers five model sizes (nano - extra-large). Training all variants on the initial dataset revealed that the large model achieved the highest segmentation mAP@0.5-0.95 score (0.744), as shown in Table 1:

Table 1. Performance for each YOLOv8-seg model size using mAP@0.5 and mAP@0.5-0.95 on the initial dataset.

Metric	Model size				
	nano (n)	small (s)	medium (m)	large (l)	extra large (x)
mAP@0.5 Segmentation	0.944	0.961	**0.967**	0.962	0.959
mAP@0.5-0.95 Segmentation	0.686	0.718	0.734	**0.744**	0.721

Following this, both medium and large versions of YOLOv8-seg were trained on an expanded dataset with a revised labeling strategy. Two training approaches were tested: (1) training from scratch and (2) fine-tuning the previously trained models on the refined dataset. Both used identical training conditions, employing the Adam optimizer (learning rate: 0.01, batch size: 16, weight decay: 0.0005) and data augmentations (random rotations, scaling, horizontal flipping, mosaic) to enhance robustness against orientation and pose variations.

Fine-tuning on the refined dataset significantly outperformed training from scratch, with the fine-tuned medium model (YOLOv8m-seg) providing the optimal balance between accuracy and computational efficiency. Figure 4 illustrates performance improvements, showing increased mAP scores and a steady decline in segmentation loss (including IoU loss), indicating effective learning and precise segmentation boundary predictions.

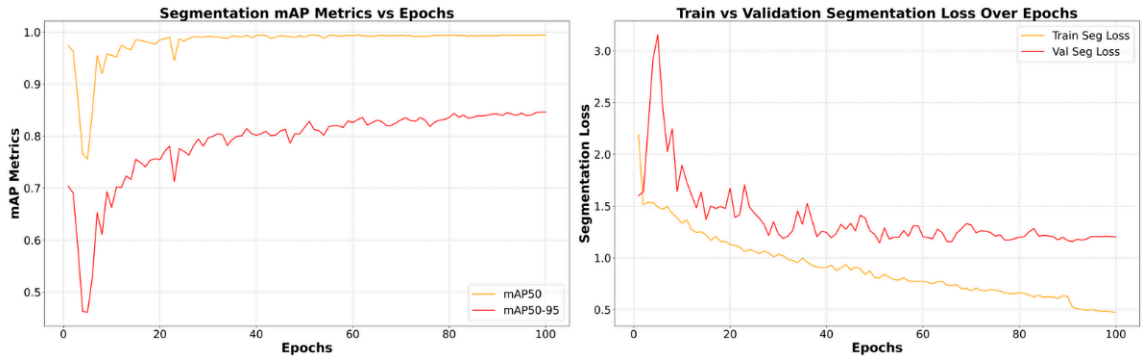

Figure 4. Training performance of the final YOLOv8m-seg model, illustrating mAP convergence (left) and decreasing segmentation loss (right), demonstrating effective learning for body part segmentation.

By adopting this fine-tuning approach, we ensured that the segmentation model performed optimally on the revised dataset.

2.3.2 Curvilinear Length Measurement

Predictions from the final YOLOv8m-seg model served as inputs for calculating the curvilinear length. The estimation process began with extracting the body centerline, which is a critical step for accurate length measurement. To define this centerline, the centers of gravity (CoGs) for the insect's head, thorax, and abdomen were computed based on the pixel-precise segmentation results. These CoGs, representing the geometric centers of each body part, were used as reference points for constructing a spatial curve along the

insect's body. A parametric cubic spline interpolation was then applied to model this curve. This method connects the CoGs using a series of cubic polynomial segments, ensuring smooth transitions and precise representation of complex body shapes. To fully capture the body length, the spline was linearly extended at both ends, linking the CoGs of the head and abdomen to the respective boundaries of each body part. The total curvilinear length was calculated by summing the lengths of all spline segments (see Appendix A in Supporting Information for details).

2.3.3 Volume Estimation

Predictions from the segmentation model were also used as inputs for classical image processing algorithms to estimate the volumes of the insect's head, thorax, and abdomen. To estimate volume, orthogonal lines were generated at equidistant intervals along the body centerline, constructed earlier from the centers of gravity (CoGs) as described in Section 2.3.2. The local slope was computed at each sampled point along the centerline (the midline), and the orthogonal line was defined as its negative reciprocal. These lines extended perpendicularly from the centerline until they intersected the boundary of the combined segmentation mask. For accurate volume computation, each orthogonal line was trimmed to match the segmented region it intersected—head, thorax, or abdomen. Each cross-section was then assigned to the appropriate body part. Assuming circular cross-sections, the diameters were used to estimate the volume of individual segments. This process ensured precise alignment of each cross-section with the segmented regions, enabling accurate length and volume estimation. A similar method was previously applied in the DiversityScanner framework (Wührl et al., 2022).

Each segment between two adjacent orthogonal lines was modeled as a frustum of a cone. This circular approximation is suitable given the generally round shape of insect body cross-sections. It offers a good balance between computational efficiency and estimation accuracy when working with 2D image data. The volume V_i^{px} can be calculated using Equation 2.1 or Equation 2.2. The variable d_i represents the diameter of the cross-section at point i along the centerline. It is measured as the distance between the two boundary points along the orthogonal line at i. The width of the truncated cone (h_i) is calculated by saving the position of the generated points on the centerline. These points correspond to the center of the respective cross-sectional area $A_i, A_{i+1}, \ldots, A_{i+n}$.

$$V_i^{px} = \frac{h_i}{12}\pi \times (d_i^2 + d_i d_{i+1} + d_{i+1}^2),\ i \in \{1, 2, \ldots, N_{midline} - 1\} \tag{2.1}$$

$$V_i^{px} = \frac{h_i}{3} \times (A_i + A_{i+1} + \sqrt{A_i \times A_{i+1}}),\ i \in \{1, 2, \ldots, N_{midline} - 1\} \tag{2.2}$$

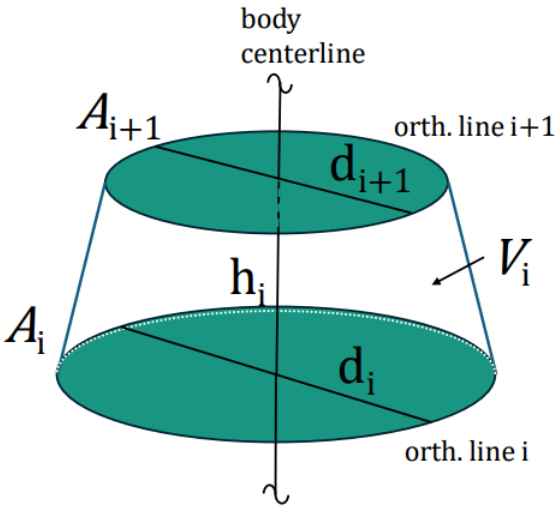

Figure 5. Principle of volume estimation using the truncated cone (frustum) method for a body segment, showing key parameters (height h_i, diameters d_i, d_{i+1}).

The total volume V_{total}^{px} of the component can be approximated by calculating the piecewise volumes and summing them (Equation 2.3).

$$V_{total}^{px} = \sum_{i=1}^{n-1} V_i^{px} \tag{2.3}$$

2.4 Supporting Information

2.4.1 Calibration and Unit Conversion

This calibration applies to both OBB length and segmentation length, as well as volume measurements. A calibration factor is applied to convert measurements from pixels to metric units (millimeters). This factor, derived using a reference object of known dimensions (a 2 mm micrometer slide divided into 200 parts), translates the pixel dimensions into real-world measurements. It is measured using the same experimental setup for capturing specimen images. After taking the picture, the number of pixels can be

measured with standard photography software. The parameter $\hat{k}$ can be calculated using the trivial formula (Equation 2.4):

$$\hat{k} = \frac{\text{Object Size in mm}}{\text{Object Size in px}} \tag{2.4}$$

To correct for resolution anisotropy, images were captured in both horizontal and vertical orientations, and the final calibration factor is the average of both measurements.

Additionally, the Entomoscope setup utilizes multiple telecentric lenses for scalability across different magnification levels. Unique calibration factors were computed for each lens configuration to maintain accuracy across magnification levels, ensuring measurement reliability and comparability across methods and with physical measurements.

Figure 6 illustrates the complete segmentation-based workflow, including the raw image, segmentation masks, centerline with CoGs, orthogonal cross-sections, and the final volume estimation, providing a clear visual summary of the process.

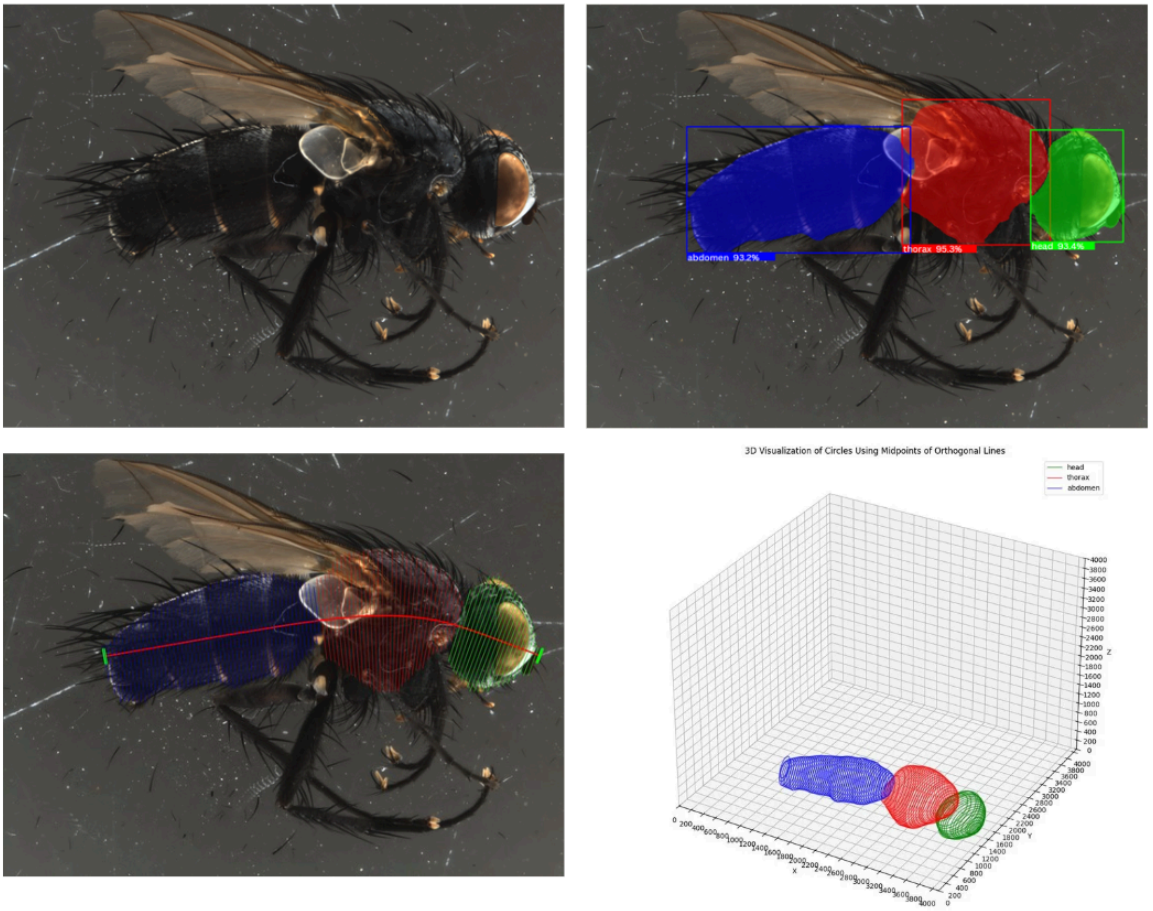

Figure 6. Workflow for segmentation-based length and volume estimation: (top left) input image, (top right) body part segmentation, (bottom left) centerline and orthogonal line extraction, (bottom right) volumetric reconstruction.

2.4.2 Software and Tools

Both methods were implemented using Python (version 3.11) and PyTorch as the primary deep learning framework, with most post-processing tasks handled using OpenCV. Each model was trained on the HAICORE high-performance computing infrastructure at the Karlsruhe Institute of Technology (KIT), utilizing GPU nodes equipped with NVIDIA A100 GPUs. While trained on HPC resources for efficiency, model inference and fine-tuning on new taxa could be feasible on standard workstations equipped with modern

GPUs. The code developed for this study, including a user-friendly application for testing both the OBB and segmentation methods, is available on GitLab at the InsectMorphoAI Repository.

3. Experimental Validation

3.1 Validation Setup and Procedures

3.1.1 Specimen Selection and Imaging

A set of 100 tachinid fly specimens belonging to 23 species was selected for detailed validation. These specimens were chosen based on the availability of high-quality lateral view images captured using the Entomoscope setup described previously (Section 2.1.1), ensuring consistent imaging conditions for applying both developed methods. While the OBB model was trained on a broad range of taxa, we utilized 100 tachinid specimens for a quantitative validation to evaluate the performance of both methods.

3.1.2 Ground Truth Measurements

3.1.2.1 Length Measurement

To establish ground truth length measurements for comparison, the same 100 specimens were imaged using a Zeiss Axio Zoom V16 photomicroscope equipped with a Plan Z 1.0x/0 objective and PI 10x/23 eyepiece. Each specimen was positioned to match the same view of the Entomoscope images. Stacked images were acquired, and the Zeiss software (version 3.2.090) was used for manual length measurements. Two types of manual measurements were performed:

1. Linear Length: A straight-line measurement between the head's anterior-most point and the abdomen's posterior-most point, used as the ground truth for the OBB method's linear length estimates.
2. Curvilinear Length: A spline curve measurement following the body's central axis, used as the ground truth for the segmentation method's curvilinear length estimates.

Care was taken to use consistent start and end points corresponding to those targeted by the automated methods to ensure direct comparability.

3.1.2.2 Biomass Proxy Measurement

To validate the segmentation method's volume estimates as a proxy for biomass, wet and dry weights of the same 100 specimens were measured. Prior to weighing, the legs were removed from the specimens to ensure consistency and focus on the main body mass comparable to the estimated volume.

For wet weighing, specimens stored in ethanol were transferred to water for one hour, then surface water was removed by gently blotting each side of the specimen on absorbent paper for five seconds (total 10 seconds), avoiding compression. Weights were measured using a Gibertini Europe 60 scale with an accuracy of 0.0001 g. For dry weighing, specimens were dehydrated at 60°C for 24 hours before weighing on a Gibertini E50S/3 Semi-Micro Scale, accurate to 0.00001 g. Weighing was conducted under controlled temperature and humidity within the shortest possible time frame after desiccation to prevent specimen rehydration. We further weighed the dried legs of specimens from selected species to assess their approximate contribution to the total weight. Replicate measurements (n=3) were performed on a subset of specimens for both wet and dry weighing to assess measurement consistency.

3.1.3 Attempted Liquid Displacement Validation

We initially tried to validate volume estimates using liquid displacement Ciborowski (1983). However, technical challenges, including micropipette calibration issues, caused inconsistent and imprecise measurements. Additionally, the required accuracy exceeded the sensitivity of this method. Thus, we chose wet and dry weighing instead, providing more reliable and strongly correlated indirect validation.

3.2 Evaluation Metrics

The following metrics were used to assess the performance of the automated methods against the ground truth measurements:

3.2.1 Error Metrics for Length

Accuracy of length estimation for both the OBB (vs. Zeiss linear) and the segmentation methods (vs. Zeiss curvilinear) was assessed using: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). MAE measures the average magnitude of the prediction errors, while RMSE emphasizes larger errors by giving them more weight. Lower values for both metrics indicate higher accuracy in length estimation.

3.2.2 Correlation Coefficients

Pearson correlation coefficients (R) and Coefficients of Determination (R^2) were calculated to quantify the linear relationship and explained variance, respectively, between estimated and measured values. A Pearson correlation coefficient (R) close to one indicates a strong positive linear correlation. The coefficient of determination (R^2)

represents the proportion of variance in the measured values that the estimated values can explain.

3.2.3 Regression Analysis for Biomass Prediction

Building upon the preliminary correlation analysis, regression techniques were employed to model the relationship between derived morphological features and measured biomass (wet and dry weight). Both linear regression and a non-linear approach were applied using eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016). Linear regression provided insights into linear trends for length and volume measurements, while the XGBoost method was used to capture more complex, non-linear relationships to predict biomass from the image-derived features.

4. Results

4.1 OBB Method: Length Estimation Performance

The OBB method provided precise estimates of linear insect length, closely matching the manual linear measurements obtained from the Zeiss photomicroscope, which serve as the point of comparison. As detailed in Table 2, this comparison showed a strong positive linear correlation (Pearson $R = 0.988$, $R^2 = 0.976$) and low measurement errors (MAE = 0.211 mm, RMSE = 0.290 mm). We also observed moderate correlations between OBB-derived length and specimen weights (Figure 7a, d), specifically with wet weight ($R = 0.793$, $R^2 = 0.630$) and dry weight ($R = 0.781$, $R^2 = 0.610$).

4.2 Segmentation Method: Performance for Length, Volume, and Biomass Prediction

The segmentation-based method provided estimates for curvilinear length and volume. The relationship between these estimates and measured biomass was evaluated.

4.2.1 Curvilinear Length Estimation

The segmentation method estimated the curvilinear body length with high accuracy relative to the manual Zeiss spline measurements (ground truth). Table 2 shows a strong correlation (Pearson R = 0.976, R^2 = 0.953) and low error metrics (MAE = 0.309 mm, RMSE = 0.408 mm). Correlations between the segmentation-derived curvilinear length and biomass were similar to those observed for OBB length (Figure 7b, e), with R = 0.811 (R^2 = 0.658) against wet weight and R = 0.792 (R^2 = 0.627) against dry weight.

Table 2. Comparison of OBB and Segmentation Length Estimation Accuracy against Zeiss Ground Truth Measurements.

Feature / Method	Comparison Target	Pearson R	R^2	MAE (mm)	RMSE (mm)
OBB linear length	Zeiss linear length	0.988	0.976	0.211	0.290
Seg curved length	Zeiss curved length	0.976	0.953	0.309	0.408

4.2.2 Volume Estimation and Correlation with Biomass

Volume estimates from the segmented body parts exhibited the strongest linear correlations with measured specimen biomass among the tested features. Figure 7 (panels c, f) visually

demonstrates these relationships and provides the quantitative metrics. The correlation with wet weight was very high (R = 0.938, R² = 0.880), as was the correlation with dry weight (R = 0.907, R² = 0.823). These results indicate that volume linearly explains a large proportion (82-88%) of this taxon's biomass variance.

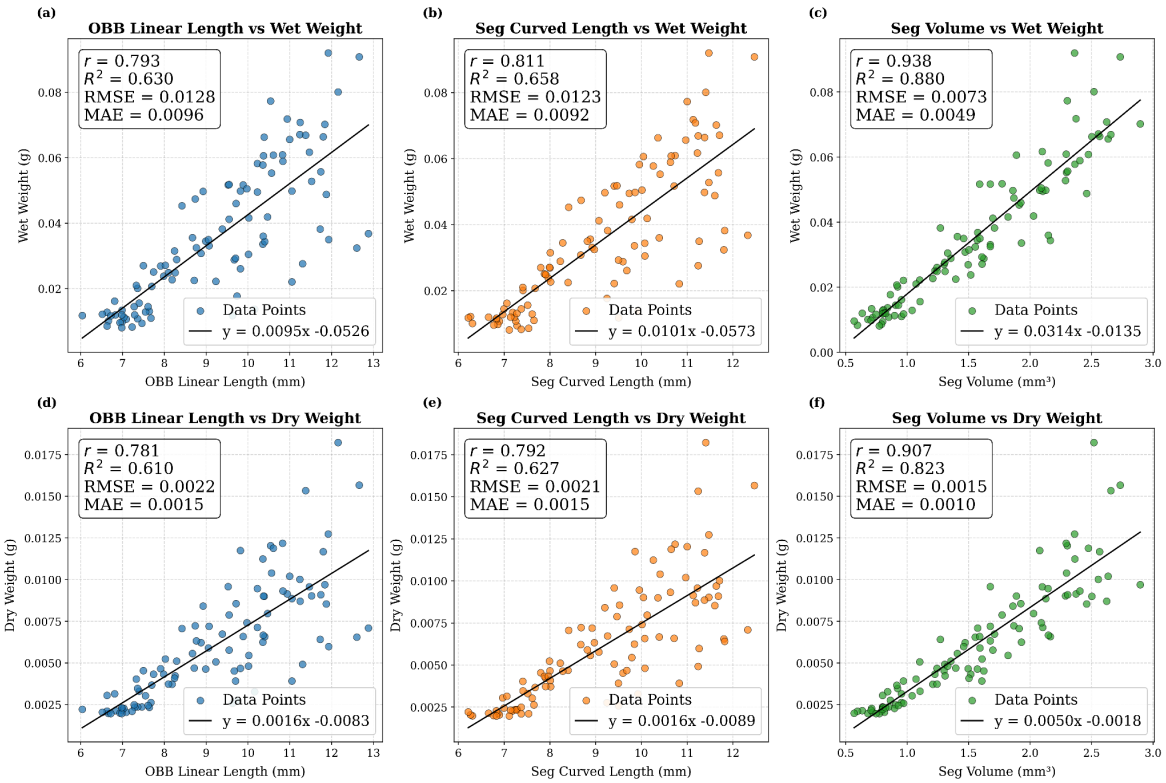

Figure 7. This figure compares the linear relationships between three image-derived features (OBB linear length, Segmentation curved length, and Segmentation volume) and measured wet weight (panels a-c) and dry weight (panels d-f) for 100 specimens. Each panel displays the data points, the linear regression fit (black line), key metrics (R, R², RMSE, MAE) for the linear prediction, and the regression line equation.

4.2.3 Biomass Prediction using Regression Models

To model the relationship between image-derived metrics and biomass, linear and non-linear regression models were employed to predict wet and dry weights using features derived from the segmentation method.

Linear Weight Prediction: The performance of simple linear regression, using each image-derived feature individually to predict biomass, is detailed in Figure 7 (panels a-f), which displays the Pearson R, R^2, MAE (g), and RMSE (g) for each relationship. Comparing these, segmentation-derived volume yielded the most accurate predictions. For wet weight prediction, volume achieved R = 0.938, R^2 = 0.880, MAE = 0.0049 g, and RMSE = 0.0073 g (Figure 7c). For dry weight prediction, it yielded R = 0.907, R^2 = 0.823, MAE = 0.0010 g, and RMSE = 0.0015 g (Figure 7f). Contextually, the contribution of legs to total dry weight was assessed in a subset of specimens in Table S3 (Supporting Information); dried legs constituted approximately 16.99% of the total dry weight on average (standard deviation 4.12%, range 11.07%-25.79%).

Non-Linear Weight Prediction with XGBoost: An eXtreme Gradient Boosting (XGBoost) analysis was conducted to model potential non-linear relationships between morphological traits and biomass. Using segmentation-derived volume and length as input features, an XGBoost model was trained and optimized using a grid search with five-fold cross-validation (hyperparameters included learning_rate, max_depth, gamma). The data were divided into 80% training and 20% testing sets, and separate models were developed for wet and dry weight predictions. The optimal parameters for the wet weight model yielded a test with an R^2 of 0.780, an MAE of 0.0082 g, and an RMSE of 0.0107 g; for the dry weight model, the test set performance included an R^2 of 0.753, an MAE of 0.0014 g, and an RMSE of 0.0019 g. Figure 8 displays 3D scatter plots showing the relationship

between the input features and the XGBoost-predicted weights for the training and test data.

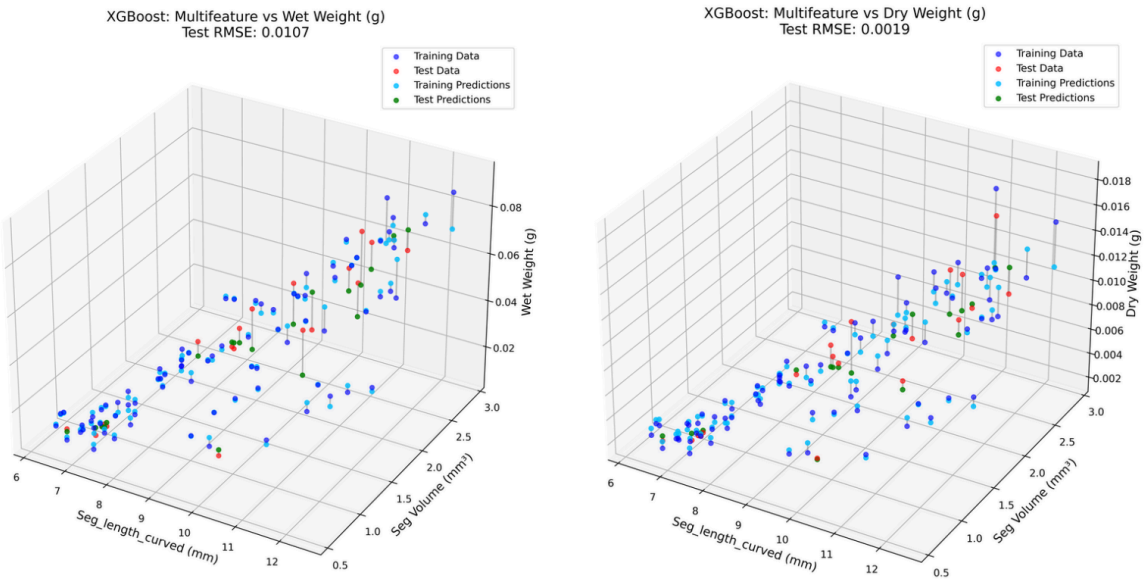

Figure 8. Non-linear XGBoost models predict wet (left) and dry (right) biomass from combined segmentation-derived length and volume features. Plots show predicted weights versus input features for the training data (blue points), test data (red points), and the model's predictions on the test data (green points).

5. Discussion

This study introduces two complementary deep learning methods for the automated estimation of insect morphometrics, body length, volume, and derived biomass estimates, designed to integrate into modern high-throughput insect analysis pipelines. We developed a general Oriented Bounding Box (OBB) method for rapid, adaptable length estimation across diverse taxa, and a more specialized instance segmentation approach for detailed, taxon-specific morphometric analysis.

The OBB method proved highly effective for rapid linear length estimation, achieving excellent accuracy (MAE = 0.211 mm, R = 0.988) compared to the manual Zeiss photomicroscope. Its computational efficiency and ability to handle varied specimen orientations without complex calibration make it a valuable tool for quickly processing large, morphologically diverse samples, such as those from Malaise traps for initial size-sorting, or population analyses in large-scale biodiversity surveys. Its reliance on overall body shape rather than fine morphological details facilitates its broad applicability, as demonstrated by its training on multiple insect orders (Diptera, Hymenoptera, Coleoptera) (Appendix B). While OBB-derived length can provide initial biomass estimates via established allometries, its primary strength lies in efficient, high-throughput length determination.

Complementing the OBB approach, the segmentation-based method delivers more granular morphometric data, including precise curvilinear length and volumetric estimates of individual body segments (head, thorax, abdomen). This method demonstrated high accuracy for curvilinear length (MAE = 0.309 mm, R = 0.976). Crucially, the strong correlations observed between segmentation-derived volume and both wet (R = 0.937) and dry (R = 0.907) weight validate its use as a reliable, non-invasive proxy for biomass. Such detailed, taxon-specific morphometric data is particularly valuable where the OBB's simpler geometry may be insufficient or where deeper biological insights are sought. For instance, precise volume measurements are essential for detailed functional trait ecology (e.g., linking body composition to metabolic rates), understanding intraspecific variation (e.g., sexual dimorphism, developmental changes), assessing population-level responses in specific pest or beneficial species, or for soil biodiversity assessment where biomass of specific decomposer groups is key. Our validation relied on wet/dry weighing, as standard

liquid displacement methods proved insufficiently precise and technically challenging for small insects, yielding inconsistent results in our attempts. The weighing protocol provided robust, reproducible measurements for comparison.

Preliminary experiments exploring non-linear modeling (XGBoost) using segmentation-derived length and volume showed that these models effectively learned relationships with biomass. However, their predictive performance on the held-out test set (Wet R^2 = 0.78, Dry R^2 = 0.75) did not substantially surpass that of the simpler linear model based on volume alone (Wet R^2 = 0.88, Dry R^2 = 0.82) for this dataset. This suggests that for Tachinidae under these conditions, volume is the dominant predictor, but non-linear approaches might offer advantages with more diverse taxonomic data or where interactions between morphological features and biomass become more significant.

Our automated, image-based methods offer significant advantages over traditional morphometric techniques, which are often invasive (requiring drying and weighing), labor-intensive, time-consuming, and can damage specimens, precluding subsequent analyses such as DNA barcoding (Hallmann et al., 2017; Lister & Garcia, 2018). These non-destructive approaches preserve specimen integrity, enabling their integration into workflows that combine morphological data with molecular analyses (Hartop et al., 2022). Compared to other deep learning approaches for biomass estimation, which may rely on area-based proxies with limited accuracy across size ranges (Ärje et al., 2020) or require extensive calibration per taxon group (Schneider et al., 2022), our two-pronged strategy offers flexibility. The OBB method provides rapid length estimates with minimal calibration, while the segmentation approach directly models volume from key body parts, offering a path to higher precision for targeted taxa.

The development of specialized, high-precision models, like our segmentation approach, is strategically valuable. While requiring taxon-specific training, this is facilitated by our modular application design, which allows users to easily test the implemented methods and enables seamless updates for adding new segmentation models for additional insect families. This modularity is particularly valuable given the typical rank-abundance distributions observed in insect communities, such as those collected with Malaise traps, where a limited number of dominant families often accounts for the majority of specimens (Srivathsan et al., 2023). Developing specialized models for just the ~20 most abundant families could enable detailed analyses for approximately 50% of all captured individuals, making the creation of family-specific models both feasible and highly impactful. Our planned integration of the rapid OBB method with the Entomoscope (Wührl et al., 2024) further extends accessibility, potentially enabling real-time morphometric estimation in the field and democratizing the technology.

Despite these strengths, some limitations must be acknowledged. Segmentation inaccuracies can occur, especially with obscured parts or unfavorable poses, although they are minimized through careful annotation strategies. Measurement variability may arise from positioning differences. Working on core body parts while excluding legs underestimates total biomass. In fact, our analysis indicates that for tachinids, legs represent ~17% of dry weight. We assume a circular cross-section for volume calculation, which is efficient for tachinids but might need refinement (e.g., using ellipsoidal models) for insects with distinctly different body shapes.

Overall, the complementary nature of the OBB and segmentation methods presented here provides a flexible and powerful toolkit for automated insect morphometrics. These

approaches offer scalable, non-invasive solutions that can significantly enhance data acquisition for a wide range of ecological and evolutionary studies, paving the way for deeper insights into insect biodiversity and function.

6. Conclusion

This study successfully developed and validated two complementary automated methods for insect morphometric analysis: an OBB-based approach providing rapid and broadly applicable length estimation, and a detailed segmentation-based method enabling precise curvilinear length and volume analysis. We demonstrated the OBB method's high accuracy for linear length and the segmentation method's capacity to estimate volume, which strongly correlates with measured biomass, thus serving as a reliable, non-invasive proxy.

By enabling integration into high-throughput biodiversity analysis pipelines, our methods significantly advance traditional approaches by offering accurate, scalable tools that preserve specimen integrity. These innovations are pivotal for modern ecological and evolutionary research, providing deeper insights into functional ecology, population dynamics, trophic interactions, ecological dynamics more broadly, and the impacts of global change. The modular and adaptable design, particularly the segmentation approach, ensures continued relevance and facilitates future extensions across diverse taxa through strategic model development for key groups.

Future research should focus on refining volume estimation algorithms, expanding training datasets taxonomically and morphologically, and integrating the OBB method with real-time imaging devices for potential field deployment. Exploring extensions of these methods to accommodate images of dried, pinned specimens also presents a promising

avenue. Developing user-friendly interfaces and cloud-based processing pipelines will further increase accessibility, ensuring these automated morphometric techniques remain valuable tools for ecological and evolutionary research.

Acknowledgments

Our work was supported by funding from: the Center for Integrative Biodiversity Discovery at the Museum für Naturkunde Berlin and by grant #ZF4717901SK9 of the program Natural, Artificial and Cognitive Information Processing (NACIP) of the Helmholtz-Association, Germany; Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition; European Union–NextGenerationEU as part of the National Biodiversity Future Center, Italian National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment 1.4 (CUP: B83C22002950007).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability

All image data used in this study are available via (Shirali & Ascenzi, 2025). The code developed for this study is available on GitLab at the InsectMorphoAI Repository**.**

References

Ärje, J., Melvad, C., Jeppesen, M. R., Madsen, S. A., Raitoharju, J., Rasmussen, M. S., Iosifidis, A., Tirronen, V., Gabbouj, M., Meissner, K., & Høye, T. T. (2020).

Automatic image-based identification and biomass estimation of invertebrates. *Methods in Ecology and Evolution*, *11*(8), 922–931. https://doi.org/10.1111/2041-210X.13428

Ascenzi, A., Wührl, L., Feng, V., Klug, N., Pylatiuk, C., Cerretti, P., & Meier, R. (n.d.). EntoSieve: Automated Size-Sorting of Insect Bulk Samples to Aid Accurate Megabarcoding and Metabarcoding. *Molecular Ecology Resources*, *n/a*(n/a), e14097. https://doi.org/10.1111/1755-0998.14097

Bazzo, C. O. G., Kamali, B., Hütt, C., Bareth, G., & Gaiser, T. (2023). A Review of Estimation Methods for Aboveground Biomass in Grasslands Using UAV. *Remote Sensing*, *15*(3), 639. https://doi.org/10.3390/rs15030639

Bosch, J., & Vicens, N. (2002). Body size as an estimator of production costs in a solitary bee. *Ecological Entomology*, *27*(2), 129–137. https://doi.org/10.1046/j.1365-2311.2002.00406.x

Büyükyilmaz, E., & Tseng, M. (2022). Developmental temperature predicts body size, flight, and pollen load in a widespread butterfly. *Ecological Entomology*, *47*(5), 872–882. https://doi.org/10.1111/een.13177

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., Wardle, D. A., Kinzig, A. P., Daily, G. C., Loreau, M., Grace, J. B., Larigauderie, A., Srivastava, D. S., & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*(7401), 59–67. https://doi.org/10.1038/nature11148

Caruso, V., Shirali, H., Bouget, C., Cerretti, P., Curletti, G., Groot, M. de, Groznik, E., Gutowski, J. M., Pyliatuk, C., Roques, A., Sallé, A., Sweeney, J., Wührl, L., & Rassati, D. (2025). Image-based recognition using advanced neural networks can

aid surveillance of Agrilus (Coleoptera, Buprestidae) jewel beetles. *ARPHA Preprints*, *6*, e154842. https://doi.org/10.3897/arphapreprints.e154842

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Ciborowski, J. J. H. (1983). A SIMPLE VOLUMETRIC INSTRUMENT TO ESTIMATE BIOMASS OF FLUID-PRESERVED INVERTEBRATES. *The Canadian Entomologist*, *115*(4), 427–430. https://doi.org/10.4039/Ent115427-4

Ganihar, S. R. (1997). Biomass estimates of terrestrial arthropods based on body length. *Journal of Biosciences*, *22*(2), 219–224. https://doi.org/10.1007/BF02704734

Habel, J. C., Ulrich, W., Biburger, N., Seibold, S., & Schmitt, T. (2019). Agricultural intensification drives butterfly decline. *Insect Conservation and Diversity*, *12*(4), 289–295. https://doi.org/10.1111/icad.12343

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., Stenmans, W., Müller, A., Sumser, H., Hörren, T., Goulson, D., & Kroon, H. de. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLOS ONE*, *12*(10), e0185809. https://doi.org/10.1371/journal.pone.0185809

Hartop, E., Lee, L., Srivathsan, A., Jones, M., Peña-Aguilera, P., Ovaskainen, O., Roslin, T., & Meier, R. (2024). Resolving biology's dark matter: Species richness, spatiotemporal distribution, and community composition of a dark taxon. *BMC Biology*, *22*(1), 215. https://doi.org/10.1186/s12915-024-02010-z

Hartop, E., Srivathsan, A., Ronquist, F., & Meier, R. (2022). Towards Large-Scale Integrative Taxonomy (LIT): Resolving the Data Conundrum for Dark Taxa. *Systematic Biology*, *71*(6), 1404–1422. https://doi.org/10.1093/sysbio/syac033

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018). *Mask R-CNN* (No. arXiv:1703.06870). arXiv. https://doi.org/10.48550/arXiv.1703.06870

Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_1), S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Hechinger, R. F. (2013). A Metabolic and Body-Size Scaling Framework for Parasite Within-Host Abundance, Biomass, and Energy Flux. *The American Naturalist*, *182*(2), 234–248. https://doi.org/10.1086/670820

Helicon Soft Ltd. (2025) Helicon Focus (version 8). Available at: https://www.heliconsoft.com/heliconsoft-products/helicon-focus/ (Accessed: 12 May 2025).

Hereld, M., Ferrier, N. J., Agarwal, N., & Sierwald, P. (2017). Designing a High-Throughput Pipeline for Digitizing Pinned Insects. *2017 IEEE 13th International Conference on E-Science (e-Science)*, 542–550. https://doi.org/10.1109/eScience.2017.88

HumanSignal/label-studio. (2024). [JavaScript]. HumanSignal. https://github.com/HumanSignal/label-studio (Original work published 2019)

Jahant-Miller, C., Miller, R., & Parry, D. (2022). Size-dependent flight capacity and propensity in a range-expanding invasive insect. *Insect Science*, *29*(3), 879–888. https://doi.org/10.1111/1744-7917.12950

Jocher, G., Chaurasia, A., & Qiu, J. (2023). *Ultralytics YOLO* (Version 8.0.0) [Python]. https://github.com/ultralytics/ultralytics (Original work published 2022)

Karlsson, D., Forshage, M., Holston, K., & Ronquist, F. (2020). The data of the Swedish

Malaise Trap Project, a countrywide inventory of Sweden's insect fauna. *Biodiversity Data Journal*, *8*, e56286. https://doi.org/10.3897/BDJ.8.e56286

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything* (No. arXiv:2304.02643). arXiv. https://doi.org/10.48550/arXiv.2304.02643

Klug, N., Kramer, M., Mazrek, F., Wührl, L., Shirali, H., Meier, R., & Pylatiuk, C. (2024). *Automated Photogrammetric Close-Range Imaging System for Small Invertebrates Using Acoustic Levitation*. https://doi.org/10.36227/techrxiv.172651022.21831566/v1

Knyshov, A., Hoang, S., & Weirauch, C. (2021). Pretrained Convolutional Neural Networks Perform Well in a Challenging Test Case: Identification of Plant Bugs (Hemiptera: Miridae) Using a Small Number of Training Images. *Insect Systematics and Diversity*, *5*(2), 3. https://doi.org/10.1093/isd/ixab004

LaBarbera, M. (1989). Analyzing Body Size as a Factor in Ecology and Evolution. *Annual Review of Ecology, Evolution, and Systematics*, *20*(Volume 20, 1989), 97–117. https://doi.org/10.1146/annurev.es.20.110189.000525

Lister, B. C., & Garcia, A. (2018). Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proceedings of the National Academy of Sciences*, *115*(44), E10397–E10406. https://doi.org/10.1073/pnas.1722477115

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2023). *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection* (No. arXiv:2303.05499). arXiv. https://doi.org/10.48550/arXiv.2303.05499

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., Hooper, D. U., Huston, M. A., Raffaelli, D., Schmid, B., Tilman, D., & Wardle, D. A. (2001). Biodiversity and Ecosystem Functioning: Current Knowledge and Future Challenges. *Science*, *294*(5543), 804–808. https://doi.org/10.1126/science.1064088

Maurey, E., Marrec, R., Brusse, T., Le Provost, G., Le Roux, V., Bergerot, B., & Caro, G. (2025). When size matters: A morphological measurement that informs on the potential pest control function by soil arthropod communities. *Journal of Pest Science*. https://doi.org/10.1007/s10340-025-01879-1

Page, R. D. M. (2016). DNA barcoding and taxonomy: Dark taxa and dark texts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150334. https://doi.org/10.1098/rstb.2015.0334

Pellegrino, N., Gharaee, Z., & Fieguth, P. (2022). *Machine Learning Challenges of Biological Factors in Insect Image Data* (No. arXiv:2211.02537). arXiv. https://doi.org/10.48550/arXiv.2211.02537

Penell, A., Raub, F., & Höfer, H. (2018). Estimating biomass from body size of European spiders based on regression models. *The Journal of Arachnology*, *46*(3), 413–419.

Plato, R. (2010). *Numerische Mathematik kompakt*. Vieweg+Teubner. https://doi.org/10.1007/978-3-8348-9644-5

Poulin, R., & George-Nascimento, M. (2007). The scaling of total parasite biomass with host body mass. *International Journal for Parasitology*, *37*(3), 359–364. https://doi.org/10.1016/j.ijpara.2006.11.009

Rodríguez-Lozano, B., Rodríguez-Caballero, E., Maggioli, L., & Cantón, Y. (2021). Non-Destructive Biomass Estimation in Mediterranean Alpha Steppes: Improving Traditional Methods for Measuring Dry and Green Fractions by Combining

Proximal Remote Sensing Tools. *Remote Sensing*, *13*(15), 2970. https://doi.org/10.3390/rs13152970

Rogers, L. E., Hinds, W. T., & Buschbom, R. L. (1976). A General Weight vs. Length Relationship for Insects1. *Annals of the Entomological Society of America*, *69*(2), 387–389. https://doi.org/10.1093/aesa/69.2.387

Sample, B. E., Cooper, R. J., Greer, R. D., & Whitmore, R. C. (1993). Estimation of Insect Biomass by Length and Width. *The American Midland Naturalist*, *129*(2), 234–240. https://doi.org/10.2307/2426503

Sánchez-Bayo, F., & Wyckhuys, K. A. G. (2019). Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation*, *232*, 8–27. https://doi.org/10.1016/j.biocon.2019.01.020

Schneider, S., Taylor, G. W., Kremer, S. C., Burgess, P., McGroarty, J., Mitsui, K., Zhuang, A., deWaard, J. R., & Fryxell, J. M. (2022). Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods in Ecology and Evolution*, *13*(2), 346–357. https://doi.org/10.1111/2041-210X.13769

Schowalter, T. D. (2016). *Insect ecology: An ecosystem approach. Academic Press.*

Shirali, H., & Ascenzi, A. (2025). *Annotated Insect Image Dataset for Training Oriented Bounding Box and Segmentation Models in Morphometric Analysis* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.15483553

Shirali, H., Hübner, J., Both, R., Raupach, M., Reischl, M., Schmidt, S., & Pylatiuk, C. (2024). Image-based recognition of parasitoid wasps using advanced neural networks. *Invertebrate Systematics*, *38*(6). https://doi.org/10.1071/IS24011

Smock, L. A. (1980). Relationships between body size and biomass of aquatic insects. *Freshwater Biology*, *10*(4), 375–383.

https://doi.org/10.1111/j.1365-2427.1980.tb01211.x

Srivathsan, A., Ang, Y., Heraty, J. M., Hwang, W. S., Jusoh, W. F. A., Kutty, S. N., Puniamoorthy, J., Yeo, D., Roslin, T., & Meier, R. (2023). Convergence of dominance and neglect in flying insect diversity. *Nature Ecology & Evolution*, *7*(7), 1012–1021. https://doi.org/10.1038/s41559-023-02066-0

Srivathsan, A., Hartop, E., Puniamoorthy, J., Lee, W. T., Kutty, S. N., Kurina, O., & Meier, R. (2019). Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. *BMC Biology*, *17*(1), 96. https://doi.org/10.1186/s12915-019-0706-9

Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S. N., Wong, J., Yeo, D., & Meier, R. (2021). ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC Biology*, *19*(1), 217. https://doi.org/10.1186/s12915-021-01141-x

Tilman, D., Reich, P. B., & Knops, J. M. H. (2006). Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature*, *441*(7093), 629–632. https://doi.org/10.1038/nature04742

van Klink, R., Bowler, D. E., Gongalsky, K. B., Swengel, A. B., Gentile, A., & Chase, J. M. (2020). Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. *Science*, *368*(6489), 417–420. https://doi.org/10.1126/science.aax9931

Wada, K. (2024). *Labelme: Image Polygonal Annotation with Python* [Python]. https://doi.org/10.5281/zenodo.5711226

Wägele, J. W., Bodesheim, P., Bourlat, S. J., Denzler, J., Diepenbroek, M., Fonseca, V., Frommolt, K.-H., Geiger, M. F., Gemeinholzer, B., Glöckner, F. O., Haucke, T.,

Kirse, A., Kölpin, A., Kostadinov, I., Kühl, H. S., Kurth, F., Lasseck, M., Liedke, S., Losch, F., … Wildermann, S. (2022). Towards a multisensor station for automated biodiversity monitoring. *Basic and Applied Ecology*, *59*, 105–138. https://doi.org/10.1016/j.baae.2022.01.003

Wang, W. Y., Srivathsan, A., Foo, M., Yamane, S. K., & Meier, R. (2018). Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. *Molecular Ecology Resources*, *18*(3), 490–501. https://doi.org/10.1111/1755-0998.12751

Wührl, L., Pylatiuk, C., Giersch, M., Lapp, F., von Rintelen, T., Balke, M., Schmidt, S., Cerretti, P., & Meier, R. (2022). DiversityScanner: Robotic handling of small invertebrates with machine learning methods. *Molecular Ecology Resources*, *22*(4), 1626–1638. https://doi.org/10.1111/1755-0998.13567

Wührl, L., Rettenberger, L., Meier, R., Hartop, E., Graf, J., & Pylatiuk, C. (2024). Entomoscope: An Open-Source Photomicroscope for Biodiversity Discovery. *IEEE Access*, *12*, 11785–11794. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3355272

Yeo, D., Srivathsan, A., Puniamoorthy, J., Maosheng, F., Grootaert, P., Chan, L., Guénard, B., Damken, C., Wahab, R. A., Yuchen, A., & Meier, R. (2021). Mangroves are an overlooked hotspot of insect diversity despite low plant diversity. *BMC Biology*, *19*(1), 202. https://doi.org/10.1186/s12915-021-01088-z

Appendices

Appendix A: Cubic Spline Interpolation with Three Points

Calculating the center of gravity of the body parts leads to the three points (x_h, y_h), (x_{th}, y_{th}), and (x_a, y_a). It is assumed that a parameter t linearly interpolates these points with $t = 0$ corresponding to (x_h, y_h) and $t = 1$ corresponding to (x_a, y_a). Because of the separate interpolation with respect to x and y, initially, only the interpolation with y is presented. The process for x is analogous. The cubic spline interpolation can be formulated as follows: (Plato, 2010)

Two cubic polynomials for the intervals $[0, t_{th}]$ and $[t_{th}, 1]$ are defined. For the two intervals, the cubic polynomials are given by

$$P_{1,y}(t) = a_1 t^3 + b_1 t^2 + c_1 t + d_1$$

$$P_{2,y}(t) = a_2 (t - t_{th})^3 + b_2 (t - t_{th})^2 + c_2 (t - t_{th}) + d_2$$

The coefficients of these polynomials are determined by enforcing continuity and boundary conditions. Firstly, the polynomials must pass through the endpoints, which leads to

$$P_{1,y}(0) = y_h \implies d_1 = y_h,$$

$$P_{2,y}(1) = y_a$$

Moreover, both polynomials must interpolate the point (x_{th}, y_{th}) at t_{th}, leading to the conditions

$$P_{1,y}(t_{th}) = y_{th} \quad \text{and} \quad P_{2,y}(t_{th}) = y_{th}.$$

To ensure a smooth transition between the two polynomial segments, continuity is imposed on the first and second derivatives at t_{th}. This results in the conditions

$$P'_{1,y}(t_{th}) = P'_{2,y}(t_{th}) \quad \text{and} \quad P''_{1,y}(t_{th}) = P''_{2,y}(t_{th}).$$

Finally, for natural bounding conditions for the splines, the second derivative at the

endpoints is set to zero, yielding

$$P''_{1,y}(0) = 0 \quad \text{and} \quad P''_{2,y}(1) = 0.$$

Solving this system of linear equations gives the coefficients $a_1, b_1, c_1, a_2, b_2, c_2,$ and d_2. The complete spline curve is then a combination of $P_{1,y}(t)$ on the interval $[0, t_{th}]$ and $P_{2,y}(t)$ on $[t_{th}, 1]$, ensuring a smooth curve through the given points. For the x-coordinate interpolation, cubic spline interpolation is independently applied using two cubic polynomials, $P_{1,x}(t)$ and $P_{2,x}(t)$, corresponding to the intervals $[0, t_{th}]$ and $[t_{th}, 1]$, respectively. This interpolation process for the x-coordinates is independent of the y -coordinate interpolation but follows the same underlying principles. The complete curve is represented parametrically as $(x(t), y(t))$, where $x(t)$ is given by $P_{1,x}(t)$ or $P_{2,x}(t)$, and $y(t)$ is similarly determined by the respective spline polynomials for the y -coordinates in their corresponding intervals. The cubic spline interpolation is solved numerically. Several software packages, e.g., MATLAB and Python, provide frameworks for solving the interpolation.

Appendix B: OBB Method Application Across Diverse Arthropod Species

This appendix provides visual examples illustrating the application of the Oriented Bounding Box (OBB) method for length estimation across a diverse selection of insect species.

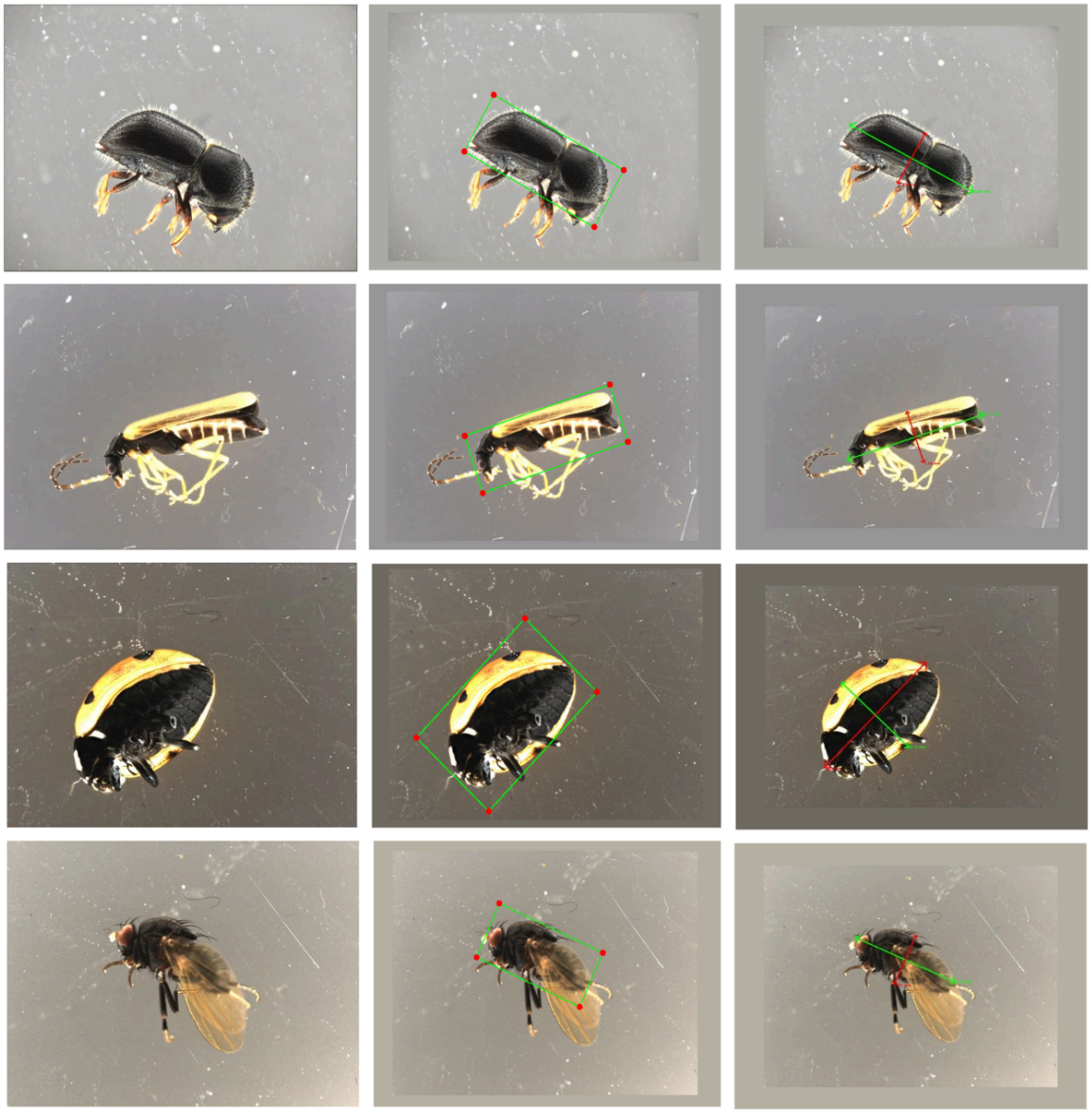

Figure S1. Oriented Bounding Box (OBB) length measurements for a selection of insect species.