Received: 4 November 2025 | Accepted: 26 February 2026

DOI: 10.1111/syen.70039

ORIGINAL ARTICLE

Systematic Entomology

Royal Entomological Society

Automated specimen triage for dark taxa: Deep learning enables orientation, sex identification and anatomical segmentation from robotic imaging

Hossein Shirali[1] | Lorenz Wührl[1] | Leshon Lee[2,3] | Nathalie Klug[1] | Rudolf Meier[2,3] | Christian Pylatiuk[1] | Emily Hartop[4]

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

[2]Center for Integrative Biodiversity Discovery, Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany

[3]Institute of Biology, Humboldt University, Berlin, Germany

[4]Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence

Hossein Shirali, Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), 76149 Karlsruhe, Germany.
Email: hossein.shirali@kit.edu

Emily Hartop, Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, NO-7491, Norway.
Email: emily.hartop@ntnu.no

Funding information

grant #ZF4717901SK9 of the program Natural, Artificial and Cognitive Information Processing (NACIP) of the Helmholtz-Association, Germany; Center for Integrative Biodiversity Discovery at the Museum für Naturkunde Berlin; Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition

Editor: Christopher Owen

Abstract

Robotic specimen processing is transforming biodiversity research by replacing manual handling with scalable systems that produce high-quality specimen images. We demonstrate that these images can be used to efficiently extract key biological information and guide targeted specimen processing by applying deep learning methods. Using a model dark taxon, Phoridae (Diptera), we show that deep learning can perform three core tasks: sex identification, determining specimen orientation and anatomical segmentation. Sex identification allows selective retention of diagnostically informative specimens, avoiding wasted effort on non-diagnostic individuals. Orientation classification enables photos of specimens with the desired orientation to be processed immediately, while suboptimally oriented specimens can be repositioned. Anatomical segmentation enables targeted processing of specimen photos that show diagnostic features. Comparative analysis of model architectures shows task-specific selection is crucial: a Convolutional Neural Network (CNN) achieved an accuracy of 0.94 for orientation, a Vision Transformer achieved 0.88 for sex and a U-Net precisely segmented nine anatomical regions with a mean IoU of 0.78. These results demonstrate that robotic imaging combined with deep learning helps in developing a high-throughput taxonomy for dark taxa, improving efficiency and utility.

KEYWORDS

automation, computer vision, dark taxa, deep learning, insect systematics, phoridae, specimen triage

INTRODUCTION

The global decline in biodiversity has created an urgent need to accelerate the documentation of life on Earth, a task for which traditional methods are insufficient (Costello et al., 2013; Stork, 2018). Modern large-scale sampling techniques, such as Malaise traps for insects, generate enormous numbers of specimens, overwhelming the capacity of the few available taxonomic experts (Karlsson et al., 2020). This influx of material, often dominated by hyperdiverse and poorly understood groups, so-called dark taxa (Hartop et al., 2024; Page, 2016), has created a severe 'taxonomic impediment', where the rate of specimen collection far outpaces the rate of identification and description. The scale of this challenge has drawn attention beyond the scientific community: dark taxa are now being highlighted in public science communication as emblematic of the biodiversity crisis (e.g. Jones, 2025). Their overwhelming dominance in global insect samples – just 20 families account for over half of the species and specimens in Malaise trap collections worldwide (Srivathsan et al., 2023) – underscores the urgency of developing scalable solutions to process and understand this hidden diversity.

To overcome this challenge, the field is moving towards a workflow that combines robotics specimen handling and imaging, high-throughput sequencing (HTS) and artificial intelligence (AI) into a unified pipeline (Meier et al., 2024; Wägele et al., 2022). This emerging paradigm of 'taxon-omics' envisions an automated workflow where specimens are first sorted by size (Ascenzi et al., 2025) and then handled by robotic systems like the DiversityScanner (Wührl et al., 2022). This system captures one high-resolution image per specimen with an option to acquire more images using 360° systems (Wührl et al., 2023) that serve as the basis for AI-driven classification (Caruso et al., 2025; Shirali et al., 2024) and morphometric analysis (Shirali et al., 2025a; Van Dam & Štarhová Serbina, 2025). These processes producing morphological data complement modern molecular workflows, such as high-throughput DNA barcoding, where specimens are often sequenced first and morphologically validated later in a 'reverse taxonomy' approach (Hartop et al., 2022; Srivathsan et al., 2021). The ultimate goal is to create a rich 'digital voucher' for every specimen, where molecular data is linked to high-quality images and quantitative morphological data, providing a holistic record of biodiversity.

However, robotics and high-throughput imaging alone are insufficient for efficient analysis; more value comes from processing specimens in ways that help downstream steps. By extracting sex, orientation and anatomical features from images, the most suitable specimens can be efficiently prioritized for downstream analyses, maximizing the utility of the collected data. This is necessary for taxa like the Phoridae (Diptera), a hyperdiverse family of small flies that has turned into a model taxon for studying dark diversity (Hartop et al., 2022; Hartop et al., 2024). Species identification in this group relies on taxonomic keys that require specimens to be viewed from standardized angles (e.g. dorsal, lateral), are often applicable to only one sex (typically males with diagnostic terminalia) and require specific characters to be visible. Manually sorting thousands of millimetre-long flies (or their images) by orientation, sex and visible anatomical features is time-consuming and limits the scalability of automated workflows. For instance, single Malaise trap samples from Uganda and Sweden have been found to contain thousands of phorid specimens representing over 600 species each (Hartop et al., 2022; Srivathsan et al., 2019), making manual processing of such samples unfeasible.

While recent computer vision studies have successfully automated single tasks like sex identification in medically important insects (Fraiwan et al., 2025; Kittichai et al., 2021; Tuda & Luna-Maldonado, 2020) or anatomical segmentation in model organisms (Le et al., 2020; Toulkeridou et al., 2023) or species descriptions (Van Dam & Štarhová Serbina, 2025), these efforts have not yet been integrated. A significant gap remains in combining these components into a comprehensive pre-processing workflow suitable for the dark taxa that dominate biodiversity samples.

Here, we address this challenge by developing a foundational AI module to automate prioritization for subsequent processing and analysis. Using Phoridae as a model system, our workflow performs three tasks on 2D specimen images:

1. Orientation Classification: Determines if the image of a specimen is in a dorsal, ventral, left lateral or right lateral view.
2. Sex Classification: Identifies if the specimen is male, female or undetermined.
3. Body Part Segmentation: Delineates nine key anatomical regions (e.g. head, thorax, terminalia).

These tasks enable specimens to be efficiently organized for downstream processing: orientation allows specimens with images in the desired position to be examined using taxonomic keys and standardized comparison, while others can be repositioned or selectively retained for subsequent steps, while sex identification prioritizes specimens informative for species-level taxonomy. Segmentation provides a foundation for future automated morphometrics, allowing the extraction of quantitative trait data directly from images. In this paper, we detail the workflow's architecture, provide a comprehensive comparison of competing deep learning models and validate its performance. We show that this automated approach is fast, accurate and shows a capacity for handling ambiguous cases in a reliable manner, offering a powerful tool to accelerate the integration of morphology into next-generation biodiversity discovery.

MATERIALS AND METHODS

Image acquisition

The dataset consisted of 1281 specimens of scuttle flies (Diptera: Phoridae) sourced from Malaise trap samples. Specimens were processed and imaged using two systems: the DiversityScanner, a robotic system for automated handling and imaging (Wührl et al., 2022) and the Entomoscope, an open-source photomicroscope (Wührl et al., 2024). Each specimen was individually imaged while submerged in ethanol. The system captured a focal stack of a minimum of five images at different depths, which were then fused into a single, high-contrast, all-in-focus composite image using Helicon Focus software (Helicon Soft, 2025).

A preprocessing step was applied to all images (Figure 1) to standardize the input for our models and remove uninformative background. A custom-trained object detection model, based on YOLOv8 architecture (Jocher et al., 2023), was used to automatically identify

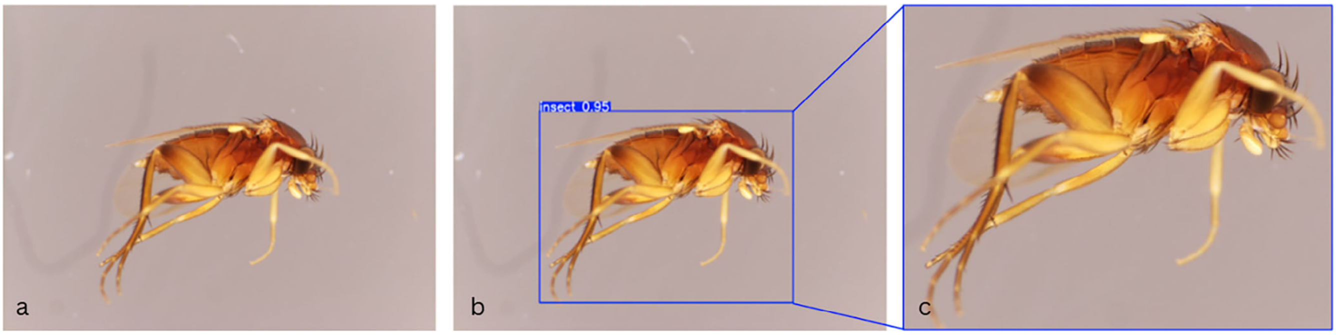

FIGURE 1 The automated image pre-processing pipeline. (a) A focus-stacked image of a Phoridae specimen. (b) A YOLOv8 model detects the insect and places a bounding box around the region of interest. (c) The final cropped image serves as the standardized input for all classification and segmentation models.

TABLE 1 Image count across tasks, classes and data subsets (training, validation and testing).

Task	Class	Training	Validation	Test
Sex classification	Male	590	126	127
	Female	173	37	38
	Undetermined	133	29	28
Total (sex classification)		896	192	193
Orientation classification	Dorsal	47	10	10
	Left Lateral	344	73	74
	Right Lateral	310	67	66
	Ventral	162	35	35
Total (orientation classification)		863	185	185
Body part segmentation		735	92	93

the primary region of interest (ROI) containing the insect. The images were then cropped to this ROI, ensuring consistent framing and centring of the specimen for all subsequent tasks. The complete, annotated image dataset is publicly available at Shirali et al. (2025c).

Ground-truth annotation

All images were annotated by taxonomic experts to create the ground-truth data for supervised learning. Annotations were performed using KAIDA, a specialized software tool for complex biological image annotation (Schilling et al., 2022). Three distinct annotation tasks were completed for each image:

- Sex Classification: Specimens were labelled as ‘Male’, ‘Female’ or ‘Undetermined’. The ‘Undetermined’ label was used when diagnostic features, primarily the terminalia (genitalia), were obscured by other body parts (e.g. legs, wings), out of focus or otherwise not clearly visible to the expert.
- Orientation Classification: Specimens were assigned to one of four positional classes: ‘Dorsal’, ‘Ventral’, ‘Left Lateral’ or ‘Right Lateral’, based on the primary view presented in the image.
- Body Part Segmentation: Nine key anatomical regions were manually delineated with pixel-wise semantic masks: Head, Thorax, Abdomen, Antenna, Palps + labella, Legs, Wings, Scutellum and Genitalia. These parts were chosen for their functional and diagnostic importance in Phoridae taxonomy.

The full dataset was split into training (70%), validation (15%) and test (15%) subsets. A stratified sampling approach was used to ensure that the proportional representation of each class within each task was maintained across the splits. This is particularly important for handling the inherent class imbalances in the dataset, such as the low number of specimens in a dorsal view. To prevent data leakage and ensure a fair evaluation, all images from a single specimen were kept within the same data split. The final distribution of images for each task is detailed in Table 1.

Deep learning architectures

To identify the optimal models for our workflow, we evaluated a range of deep learning architectures for the classification and segmentation tasks.

Classification models

For sex and orientation classification, we evaluated two major classes of deep learning models to compare their distinct architectural strategies. Convolutional Neural Networks (CNNs) (Lecun et al., 1998) are designed to detect local patterns (e.g. edges, textures) by applying filters that scan across an image. In contrast, Vision Transformers (ViTs) (Dosovitskiy et al., 2021) take a more holistic approach, treating an image as a sequence of patches and using attention mechanisms to model the relationships between them, allowing the capture of global context. Our comparison aimed to determine whether local feature extraction (CNNs) or global context modelling (ViTs) was more effective for our specific biological classification tasks. We tested architectures chosen to represent a broad spectrum of modern computer vision approaches:

- Lightweight and Efficient CNNs: We included YOLOv8 and YOLOv11 (Jocher et al., 2023) as they are state-of-the-art models known for their balance of speed and accuracy, representing a practical choice for real-world deployment. For both architectures, we systematically trained all five standard model variants (from nano to x-large) and report the results for the best-performing 'x-large' variant for brevity.
- High-Performance CNNs: To explore the upper limits of CNN performance, we included ConvNeXt-XLarge (Liu et al., 2022) and EfficientNetV2-L (Tan & Le, 2019), which are larger, more complex architectures designed for maximum accuracy. Additionally, to provide a standard CNN baseline for fair comparison with hierarchical transformers, we included ResNet50v2 (He et al., 2016a).
- Vision Transformers (ViTs): To test a fundamentally different approach, we included BEiTv2-large (Peng et al., 2022) and EVA-02-large (Fang et al., 2024). We also included Swin-Tiny (Liu et al., 2021), a hierarchical Vision Transformer that, like CNNs, computes self-attention within local windows, enabling a direct comparison of local versus global feature extraction strategies. These models employ attention mechanisms rather than local convolutions, enabling them to capture broader contextual relationships within the image.

Segmentation model

For body part segmentation, we used the U-Net architecture, a standard for biomedical image segmentation due to its robust performance and efficient encoder-decoder structure with skip connections (Ronneberger et al., 2015). To optimize feature extraction within the U-Net, we experimented with two different backbones: EfficientNet-B0, known for its balance of efficiency and accuracy, and ResNet-18 (He et al., 2016b), a well-established residual network.

Model training and implementation

All models were trained on the HAICORE high-performance computing cluster at the Karlsruhe Institute of Technology (KIT), using nodes equipped with NVIDIA A100 80GB GPUs. We used transfer learning for all models, initializing them with weights pre-trained on the ImageNet dataset (Deng et al., 2009) to accelerate training and improve generalization. The models were then fine-tuned on our Phoridae dataset using a comprehensive optimization strategy. Input image sizes were tailored to each model's architecture, typically ranging from 224 × 224 to 640 × 640 pixels, to optimize performance. We applied a suite of data augmentation techniques to increase model robustness to variations in imaging conditions. These included photometric augmentations such as random changes in brightness, contrast and saturation. For the segmentation and sex classification tasks, we also applied geometric augmentations, including horizontal flipping, scaling and rotation. For the orientation classification task, geometric augmentations that would alter the orientation label (e.g. flipping and rotation) were explicitly excluded to avoid label ambiguity. Similarly, we avoided synthetic augmentation methods (e.g. MixUp, CutMix) to preserve the specimens' geometric integrity and the visibility of diagnostic features.

Our optimization strategy focused on achieving robust generalization and preventing overfitting. We used the AdamW optimizer for all classification models with an initial learning rate of 1e-4, and the Adam optimizer for segmentation with a learning rate of 1e-3. Our training strategy had two phases. First, initial screening of all architectures using standard class weighting (calculated inversely proportional to class frequency) identified top-performing candidates. Second, for the challenging orientation task—where the 'Dorsal' class was severely underrepresented—we performed a targeted optimization using Weighted Random Sampling (oversampling). This technique was applied to our primary candidate (YOLOv8) and two representative benchmarks (ResNet50v2 and Swin-Tiny) to evaluate architectural differences under balanced data conditions. For segmentation, we used a combined loss function of Focal Loss and Dice Loss, which is effective for handling extreme imbalances between foreground and background pixels. Furthermore, we implemented dropout regularization with a rate of 0.3 during training and employed an early stopping mechanism that halted training if validation loss did not improve for 15 consecutive epochs. Models were trained for a maximum of 150 epochs for classification and 250 epochs for segmentation. All models converged successfully (see Figures S1 and S4 for training and validation curves for the best-performing models).

Evaluation

Model performance was assessed using a range of standard metrics on the held-out test set.

- Classification: We evaluated classification performance using Accuracy, Precision, Recall and F1-Score. The F1-score, which is the harmonic mean of precision and recall, is particularly useful for datasets with class imbalance.
- Segmentation: Segmentation quality was measured using the Intersection over Union (IoU, or Jaccard index) and the Dice Score

TABLE 2 Model performance on classification tasks.

Architecture	Sex accuracy	Sex F1-score	Orientation accuracy	Orientation F1-score
YOLOv8x-cls	0.88	0.82	**0.94**	**0.89**
YOLOv8x-cls*	-	-	**0.94**	**0.92**
YOLO11x-cls	0.82	0.75	0.90	0.86
BEiTv2	**0.88**	**0.85**	0.61	0.72
EVA-02	0.85	0.81	0.61	0.66
ConvNeXtXLarge	0.85	0.81	0.66	0.73
EfficientNetV2L	0.82	0.75	0.91	0.88
ResNet50v2*	-	-	0.92	0.87
Swin-Tiny*	-	-	0.91	0.89

Note: Comparison of overall accuracy and weighted F1-score for all different model architectures on the test dataset. The best-performing model for each task is highlighted in bold. Models with an asterisk (*) were retrained using Weighted Random Sampling to address class imbalance in the orientation task. Scores are colour-coded to indicate performance levels: Green (≥0.90), yellow (0.80–80.9), orange (0.70–0.79) and red (<0.70).
*Models trained with Weighted Random Sampling (oversampling) to evaluate architectural performance under balanced data conditions for the orientation task.

(an F1-score equivalent for segmentation). These metrics quantify the overlap between the predicted segmentation mask and the ground-truth annotation.

- Independent Validation: To benchmark our model against human-level performance, an independent expert taxonomist (not involved in the initial annotation) re-classified all images in the test set for both sex and orientation. This provided a direct comparison between the model and expert accuracy. Additionally, to analyse the effect of feature visibility on segmentation performance, the independent tester annotated each target body part in the test images as 'clearly visible', 'partially visible' or 'not visible' based on its photographic clarity and focus. A detailed summary of the independent validation is provided in the Supporting Information.

RESULTS

Model performance in classification tasks

Our comparative analysis of six architectures showed that the optimal model choice depended on the classification task, with sometimes CNNs and other times transformers excelling (Table 2). We thus assessed the performance gap between CNNs and Transformers further by studying whether it was driven by architecture or data imbalance. We trained two additional standard benchmarks (ResNet50v2 and Swin-Tiny) using the optimized balanced sampling strategy. For orientation classification, the balanced YOLOv8x-cls remained the top-performing model, achieving an overall accuracy of 0.94 and an improved weighted F1-score of 0.92. Interestingly, the hierarchical Vision Transformer (Swin-Tiny, F1 = 0.89) outperformed the standard CNN (ResNet50v2, F1 = 0.87). This suggests that while hierarchical feature extraction is critical for this geometric task, the specific YOLO architecture provides the best trade-off for accuracy. Notably, the global Vision Transformers (BEiTv2 and EVA-02) achieved low accuracy (0.61), confirming that global attention mechanisms are less effective than hierarchical approaches for orientation discrimination. In contrast, for sex classification, the Vision Transformer-based BEiTv2 model performed best, reaching an accuracy of 0.88 and the highest F1-score of 0.85.

A detailed analysis of the best-performing models (Table 3) shows strong performance on the most common classes (See confusion matrices in the Supporting Information: Figure S2). For orientation, the balanced YOLOv8x-cls model achieved excellent precision and recall for Left Lateral (F1 = 0.95) and Right Lateral (F1 = 0.96) views. Performance was lower for the Ventral (F1 = 0.88) and the Dorsal view (F1 = 0.89).

For sex classification, the BEiTv2 model performed well on both Male (F1 = 0.91) and Female (F1 = 0.90) classes. The F1-score for the 'Undetermined' class was lower (0.74), which is expected given the inherent ambiguity of these images.

Model performance in body part segmentation

For the body part segmentation task, the U-Net architecture with an EfficientNetB0 backbone outperformed the ResNet18 backbone (Table 4). The model trained successfully, showing stable convergence of loss and IoU scores (Figure S4) and achieved a mean IoU of 0.78 and a mean Dice score of 0.87 across all nine anatomical classes.

The model demonstrated excellent performance on large, well-defined body parts, with dice scores above 0.91 for the head, thorax, abdomen and wings. Performance was moderately lower for smaller or more complex structures like Legs (0.90), Antennae (0.84) and Genitalia (0.84). The lowest score was recorded for the Halteres (0.75), the smallest and most frequently obscured body part, which also had the lowest pixel representation in the dataset (see Figure S3 for pixel count distribution).

TABLE 3 Detailed classification report for the best-performing models.

Task	Class	Precision	Recall	F1-score	Support
Orientation	Dorsal	100	0.8	0.89	10
	Left lateral	0.96	0.93	0.95	74
	Right lateral	0.96	0.97	0.96	66
	Ventral	0.84	0.91	0.88	35
	Accuracy			0.94	185
	Macro Avg	0.94	0.9	0.92	185
	Weighted Avg	0.94	0.94	0.94	185
Sex	Female	0.94	0.87	0.9	38
	Male	0.93	0.89	0.91	127
	Undetermined	0.65	0.86	0.74	28
	Accuracy			0.88	193
	Macro Avg	0.84	0.87	0.85	193
	Weighted Avg	0.89	0.88	0.88	193

Note: Per-class precision, recall and F1-score for the balanced YOLOv8x-cls model on the orientation task and the BEiTv2 model on the sex task. Scores are colour-coded to indicate performance levels: Green (≥0.90), yellow (0.80–0.89), orange (0.70–0.79) and red (<0.70).

TABLE 4 Body part segmentation performance.

Class	IoU (EfficientNetB0)	Dice score (EfficientNetB0)	IoU (ResNet18)	Dice score (ResNet18)
Head	0.84	0.91	0.82	0.90
Thorax	0.84	0.91	0.83	0.91
Abdomen	0.85	0.92	0.83	0.91
Antenna	0.72	0.84	0.72	0.84
Palps + labella	0.71	0.83	0.72	0.84
Legs	0.82	0.90	0.81	0.89
Wings	0.88	0.93	0.87	0.93
Halteres	0.60	0.75	0.59	0.74
Genitalia	0.72	0.84	0.66	0.79
Mean	0.78	0.87	0.76	0.86

Note: Comparison of Intersection over Union (IoU) and Dice scores for the U-Net architecture with EfficientNetB0 and ResNet18 backbones on the test dataset. Scores are colour-coded to indicate performance levels: Green (≥0.90), yellow (0.80–0.89), orange (0.70–0.79) and red (<0.70).

Qualitative and independent validation

Visual inspection of the segmentation outputs shows a high degree of accuracy and regional consistency (Figure 2). The example shown in Figure 2, which is representative of the model's average performance, achieved an overall IoU of 0.79 and a Dice score of 0.87. The model's predicted masks often appear smoother and more precise than the hand-labelled ground truth, particularly around complex boundaries. This suggests that the model learns generalized representations of anatomical structures, potentially surpassing the consistency of manual annotation.

The models were also conservative. They assigned a sex to 4 of the 28 specimens that experts had labelled 'Undetermined' (Figure 3a), but the more frequent error was to classify 13 known males as 'Undetermined'. For specimens in ambiguous orientations (Figure 3b), the workflow can resolve these intermediate cases by providing both a primary and a valid secondary classification using a top-2 prediction strategy. Crucially, explainable AI techniques confirm that the models learn to focus on the correct anatomical regions; an Eigen-CAM visualization (Muhammad & Yeasin, 2020) shows the CNN-based orientation model focuses on the head and thorax (Figure 3c), while attention score maps from the Vision Transformer show that terminalia are used for sex classification (Figure 3d). The full results of the independent validation are summarized in Table 5, showing close agreement between the top models and expert performance.

DISCUSSION

We have developed and validated a multi-component deep learning workflow that extracts key biological information from specimen images, enabling targeted and efficient processing in a systematic pipeline when processing large-scale image datasets. High accuracy across orientation

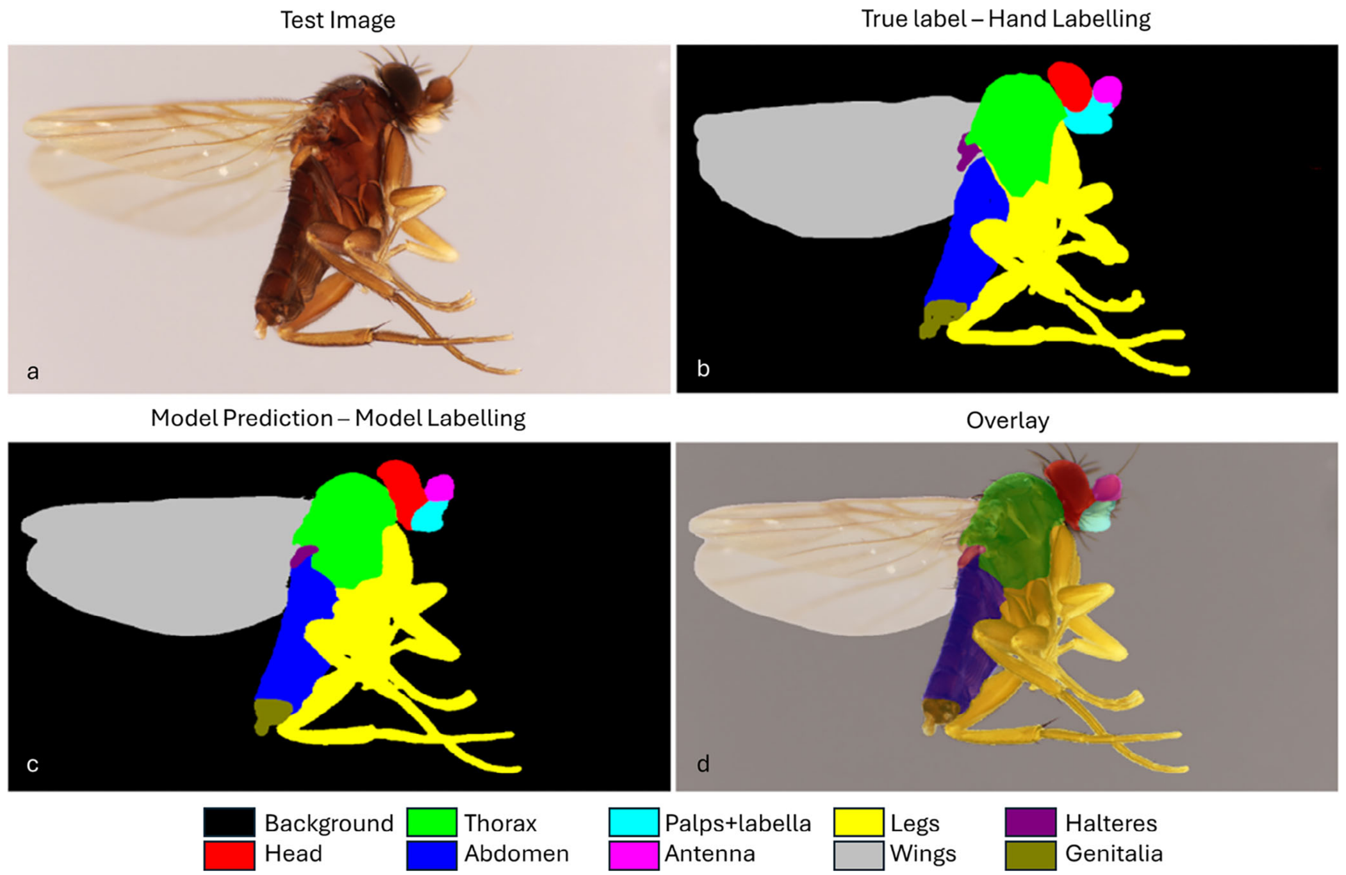

FIGURE 2 Qualitative results of body part segmentation. Comparison of (a) the original test image, (b) the hand-labelled ground-truth mask, (c) the model's predicted segmentation mask and (d) an overlay of the prediction on the original image. Each colour corresponds to a different anatomical region, as shown in the legend. The prediction for this specific example achieved an overall IoU of 0.79 and a Dice score of 0.87, reflecting the model's average performance.

classification, sex identification and anatomical segmentation demonstrates that AI can reliably prioritize informative specimens and images and thus provide a robust pre-sorting workflow for taxonomic research and trait extractions. While adaptability to other taxa remains to be tested, this framework provides a foundation for developing more efficient and scalable taxonomic workflows across diverse organismal groups. Other researchers could apply this approach to their taxa by first generating high-quality specimen images (e.g. via robotic or semi-automated imaging), annotating a representative subset for key traits (sex, orientation, anatomical regions) and training task-specific deep learning models. Importantly, previous analyses of Malaise trap samples show that a small number of taxa dominate collections—approximately ten families account for roughly 50% of all specimens (Srivathsan et al., 2023). This suggests that a limited set of well-tuned models could capture a substantial fraction of collected material, providing an efficient route to scale up taxonomic processing across diverse taxa.

An accurate and automated workflow for early-stage specimen screening

This work introduces a deep learning workflow embedded in robotic specimen processing systems to automate three critical tasks: orientation classification, sex identification and anatomical segmentation from high-resolution insect images. By enabling targeted, high-throughput prioritization of diagnostically informative specimens from hyperdiverse dark taxa, it lays the foundation for scalable, morphology-integrated biodiversity discovery. With accuracies of 0.94 for orientation and 0.88 for sex, our workflow is sufficiently accurate for practical implementation. In a high-throughput pipeline, this module could efficiently and automatically sort large samples, flagging only a small fraction for expert review. This represents a significant saving of expert time and a major step towards scaling up taxonomic research and insect species discovery.

Our comparative analysis also provided a key insight into model selection: there is no one-size-fits-all architecture. Global Vision Transformers (e.g. BEiT) struggled with orientation, while the hierarchical Swin Transformer performed competitively. This indicates that hierarchical feature extraction—whether via convolution (YOLO, ResNet) or windowed attention (Swin)—is essential for distinguishing specimen orientation. Ultimately, the YOLOv8 architecture proved superior, likely due to its effective multi-scale feature integration (Figure 3c). Conversely, the success of the ViT-based model in sex classification implies that this task requires a more holistic assessment of the image, likely integrating subtle textural cues and global morphological patterns. The use of explainable

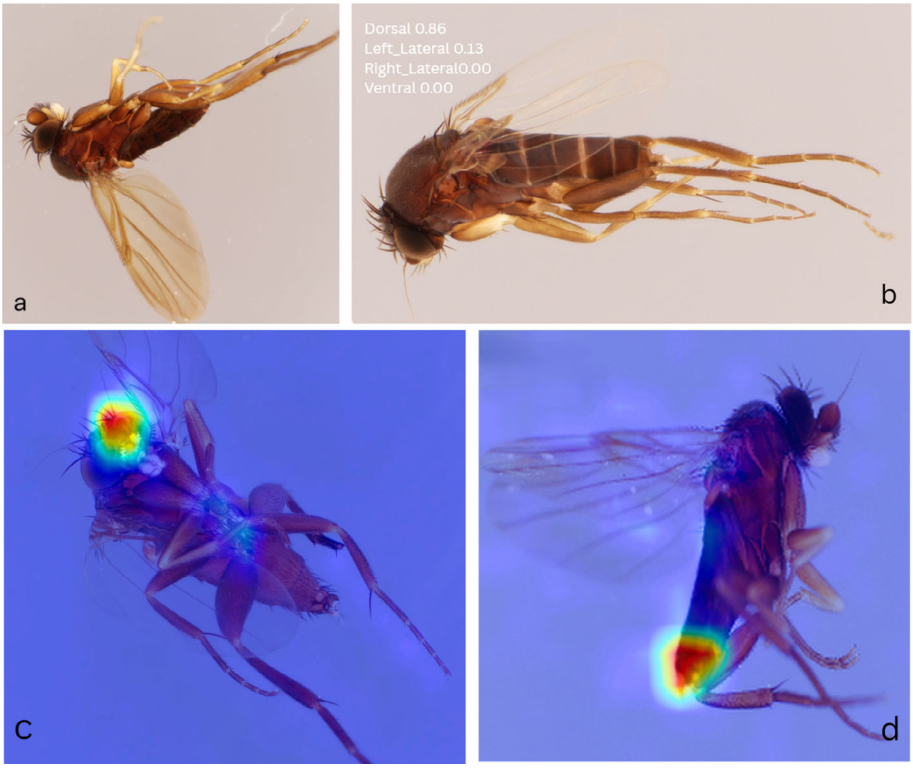

FIGURE 3 Analysis of challenging classification cases and model interpretability. (a) An example of the model assigning a sex (Male) to a specimen that experts had originally labelled as 'Undetermined'. (b) A specimen in an ambiguous orientation between dorsal and lateral views, a case that can be handled by a logically constrained top-2 prediction strategy. (c) An Eigen-CAM visualization, showing the orientation model (red heatmap), is correctly focused on the head region. (d) An attention score map from the Vision Transformer confirms that the sex classification model's attention is concentrated on the terminalia when making predictions.

TABLE 5 Comparison of expert and model predictions.

True label	Total	Prediction matches	Prediction differences	Expert's matches	Expert's differences	Prediction accuracy	Expert's accuracy
Female	38	33	5	38	0	0.87	1
Male	127	114	14	125	2	0.89	0.97
Undetermined	28	24	4	27	1	0.86	0.96
Dorsal	10	8	2	10	0	0.8	1
Left lateral	74	69	5	74	0	0.93	1
Right lateral	66	64	2	66	0	0.97	1
Ventral	35	32	3	35	0	0.91	1

Note: A summary of the agreement between the ground-truth labels, the top model predictions and the independent expert's classifications on the test set. Scores are colour-coded to indicate performance levels: Green (≥0.90), yellow (0.80–0.89), orange (0.70–0.79) and red (<0.70).

AI (Figure 3d) provides strong evidence for this, confirming that the model has learned to focus on the biologically relevant terminalia. This finding underscores the importance of empirical testing and task-specific model selection in applied AI for biology.

Beyond sorting: The foundational role of segmentation

The accurate segmentation of body parts is not an endpoint, but a gateway to a suite of downstream applications that are central to the future of digital systematics. The precise masks generated by our U-Net model can serve as the foundation for:

1. Automated Morphometrics: The pixel data from segmented regions such as the head, thorax and wings can be used to automatically calculate lengths, areas and volumes. This enables the non-invasive extraction of key functional trait data at a scale impossible with manual methods.
2. Automated Character Extraction: The workflow could be extended to automatically score discrete morphological characters crucial for

phylogenetic reconstruction and species description (e.g. presence/absence of certain bristles on the segmented thorax, or ratios of leg segments).

3. Targeted Re-imaging: In an integrated robotic pipeline, the segmentation output could direct the imaging system to perform a high-resolution scan of a specific region of interest, such as the terminalia, for detailed taxonomic study.
4. Guiding AI-based Identification: The segmentation output can act as a quality control filter for subsequent species identification models. For example, an identification model could be instructed not to process a specimen if the segmentation masks indicate that key diagnostic features, such as the terminalia, are not visible.

By providing this foundational data layer, our segmentation module is a critical technology for the next generation of quantitative and automated biodiversity analysis.

Model robustness, limitations and future directions

Our models were also conservative when faced with ambiguous evidence. An analysis of the sex classification model's performance on the 'Undetermined' class provides a key insight. While the model successfully assigned a correct sex to 4 specimens that experts had found ambiguous (e.g. Figure 3a), its dominant behaviour was to err on the side of caution. It misclassified 13 clear males as 'Undetermined', whereas it only made 6 direct misclassifications between the Male and Female classes combined. This tendency to classify uncertain cases as 'Undetermined' is a valuable safety feature in an automated sorting pipeline, as it minimizes downstream errors by flagging difficult specimens for expert review. Furthermore, we developed a method to address the ambiguity of specimens positioned between two standard orientations. By implementing a logically constrained top-2 prediction system, the workflow can flag these intermediate cases for review. This system considers a secondary prediction only if it is biologically plausible (e.g. a 'dorsal' secondary prediction for a 'lateral' primary prediction) and its confidence exceeds a 5% threshold. This provides a robust solution to a common practical challenge without forcing a single, potentially incorrect, classification. By flagging ambiguous specimens (e.g. dorsal-lateral) via inference logic, this strategy effectively serves the purpose of a multi-label framework without the need for complex re-annotation or architectural changes.

However, we acknowledge several limitations that suggest clear avenues for future work. Initially, the model's performance for orientation was limited by class imbalance (e.g. rare dorsal views). Model performance was significantly improved by implementing weighted random sampling (Dorsal F1 improved to 0.89). However, the same strategy failed to improve the success rate for sex classification (accuracy remained ∼0.88). This result suggests that the primary challenge in sex identification is not sample quantity but rather the inherent ambiguity of diagnostic features (e.g. obscured terminalia) in 'Undetermined' specimens. For segmentation, the lowest scores were for 'Halteres', a classic challenge of data imbalance. This result is further explained by our independent visibility assessment, which revealed that halteres were labelled as 'not visible' or 'partially visible' in 34% of the test images (see visibility assessment in Data S1). This confirms that the model's performance is logically constrained by the quality and visibility of features in the input data, not just by algorithmic limitations. Finally, this study was confined to a single insect family. While Phoridae represents a challenging test case, the generalizability of these models to other insect groups with different morphologies remains to be tested.

Broader implications for the future of systematics

Our study delivers a validated module for integrating deep learning into robotic biodiversity pipelines. Automating sex identification, orientation and anatomical segmentation ensures that only specimens that are diagnostically informative are carried forward for downstream molecular or morphological analyses. These capabilities are crucial for tackling the vast numbers of undescribed insect 'dark taxa' that dominate global sampling efforts. As such, workflows are adopted more broadly, and they promise to accelerate species discovery and contribute to the creation of comprehensive, high-quality digital biodiversity records.

CONCLUSION

We present a deep learning workflow that automates three core processing tasks: orientation classification, sex identification and body part segmentation. Validated on a taxonomically challenging group, this system achieves near-expert-level performance and is ready for integration into large-scale, automated biodiversity pipelines. By streamlining specimen processing triage, it marks a practical advance towards high-throughput, data-driven species discovery.

AUTHOR CONTRIBUTIONS

H.S. and E.H. conceived and designed the study. H.S. developed and trained all deep learning models, performed data analysis, created visualizations and wrote the original draft. L.W. developed the DiversityScanner and Entomoscope imaging workflows, processed specimen images and edited the manuscript. L.L. conducted ground-truth labelling, and reviewed the manuscript. N.K. optimized photogrammetric imaging protocols, supported image acquisition and contributed to manuscript preparation. R.M. conceptualized the broader biodiversity workflow, provided senior supervision, secured funding and critically revised the manuscript. C.P. conceptualized the robotic-AI integration, supervised model development, provided infrastructure support and critically revised the manuscript. E.H. conceived the study, provided taxonomic expertise on Phoridae and dark taxa, supervised data annotation, co-wrote the manuscript and oversaw project coordination. All authors reviewed, edited and approved the final manuscript.

ACKNOWLEDGEMENTS

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.

During the preparation of this work, the authors used AI tools to improve clarity and fluency in the writing process. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.This manuscript was previously made available as a preprint on bioRxiv: https://doi.org/10.1101/2025.10.02.680063. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All image data and the pre-trained model weights used in this study are available on Zenodo (https://zenodo.org/records/18429938). Code for model inference is openly available on GitLab (Shirali et al., 2025b).

ORCID

Hossein Shirali https://orcid.org/0009-0005-6884-4263

REFERENCES

Ascenzi, A., Wührl, L., Feng, V., Klug, N., Pylatiuk, C., Cerretti, P. et al. (2025) EntoSieve: automated size-sorting of insect bulk samples to aid accurate megabarcoding and metabarcoding. *Molecular Ecology Resources*, 25(6), e14097. Available from: https://doi.org/10.1111/1755-0998.14097

Caruso, V., Shirali, H., Bouget, C., Cerretti, P., Curletti, G., de Groot, M. et al. (2025) Image-based recognition using advanced neural networks can aid surveillance of agrilus (coleoptera, buprestidae) jewel beetles. [accessed 2025 Aug 7] https://doi.org/10.3897/arphapreprints.e154842

Costello, M.J., May, R.M. & Stork, N.E. (2013) Can we name earth's species before they go extinct? *Science*, 339(6118), 413–416. Available from: https://doi.org/10.1126/science.1230318

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009) ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA: IEEE, pp. 248–255. [accessed 2025 Aug 7]. https://ieeexplore.ieee.org/document/5206848. Available from: https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. et al. (2021) An image is worth 16 × 16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR 2021)*. Available from: https://doi.org/10.48550/arXiv.2010.11929

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X. & Cao, Y. (2024) EVA-02: a visual representation for neon genesis. *Image and Vision Computing*, 149, 105171. Available from: https://doi.org/10.1016/j.imavis.2024.105171

Fraiwan, M., Mukbel, R. & Kanaan, D. (2025) Using deep learning artificial intelligence for sex identification and taxonomy of sand fly species. *PLoS One*, 20(4), e0320224. Available from: https://doi.org/10.1371/journal.pone.0320224

Hartop, E., Lee, L., Srivathsan, A., Jones, M., Peña-Aguilera, P., Ovaskainen, O. et al. (2024) Resolving biology's dark matter: species richness, spatiotemporal distribution, and community composition of a dark taxon. *BMC Biology*, 22(1), 215. Available from: 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. Available from: https://doi.org/10.1093/sysbio/syac033

He, K., Zhang, X., Ren, S. & Sun, J. (2016a) Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, pp. 770–778. [accessed 2025 Oct 2]. http://ieeexplore.ieee.org/document/7780459/. Available from: https://doi.org/10.1109/CVPR.2016.90

He, K., Zhang, X., Ren, S. & Sun, J. (2016b) Identity mappings in deep residual networks. In: *European Conference on Computer Vision (ECCV)*. Cham: Springer, pp. 630–645. Available from: https://doi.org/10.1007/978-3-319-46493-0_38

Helicon Soft. (2025) Helicon Focus. Available from: https://www.heliconsoft.com/heliconsoft-products/helicon-focus/ [accessed 12 May 2025]

Jocher, G., Chaurasia, A. & Qiu, J. (2023) Ultralytics yolo. [accessed 2024 May 29]. https://github.com/ultralytics/ultralytics

Jones, B. (2025) The search for earth's most mysterious creatures is turning up extraordinary results. Vox. [accessed 2025 Oct 2]. https://www.vox.com/down-to-earth/459398/animals-species-unknown-dark-taxa

Karlsson, D., Hartop, E., Forshage, M., Jaschhof, M. & Ronquist, F. (2020) The Swedish malaise trap project: a 15 year retrospective on a countrywide insect inventory. *Biodiversity Data Journal*, 8, e47255. Available from: https://doi.org/10.3897/BDJ.8.e47255

Kittichai, V., Pengsakul, T., Chumchuen, K., Samung, Y., Sriwichai, P., Phatthamolrat, N. et al. (2021) Deep learning approaches for challenging species and gender identification of mosquito vectors. *Scientific Reports*, 11(1), 4838. Available from: https://doi.org/10.1038/s41598-021-84219-4

Le, V.-L., Beurton-Aimar, M., Zemmari, A., Marie, A. & Parisey, N. (2020) Automated landmarking for insects morphometric analysis using deep neural networks. *Ecological Informatics*, 60, 101175. Available from: https://doi.org/10.1016/j.ecoinf.2020.101175

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. Available from: https://doi.org/10.1109/5.726791

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. et al. (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, BC, Canada: IEEE, pp. 10012–10022. Available from: https://doi.org/10.1109/ICCV48922.2021.00986

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. & Xie, S. (2022) A convnet for the 2020s. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, pp. 11966–11976. [accessed 2025 Oct 2]. https://ieeexplore.ieee.org/document/9879745/. Available from: https://doi.org/10.1109/CVPR52688.2022.01167

Meier, R., Hartop, E., Pylatiuk, C. & Srivathsan, A. (2024) Towards holistic insect monitoring: species discovery, description, identification and traits for all insects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1904), 20230120. Available from: https://doi.org/10.1098/rstb.2023.0120

Muhammad, M.B. & Yeasin, M. (2020) Eigen-cam: class activation map using principal components. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, UK: IEEE, pp. 1–7. [accessed

2025 Aug 8]. https://ieeexplore.ieee.org/document/9206626. Available from: https://doi.org/10.1109/IJCNN48605.2020.9206626

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. Available from: https://doi.org/10.1098/rstb.2015.0334

Peng, Z., Dong, L., Bao, H., Ye, Q. & Wei, F. (2022) BEiT v2: masked image modeling with vector-quantized visual tokenizers. [accessed 2025 Oct 2]. http://arxiv.org/abs/2208.06366 https://doi.org/10.48550/arXiv.2208.06366

Ronneberger, O., Fischer, P. & Brox, T. (2015) U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M. & Frangi, A.F. (Eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, pp. 234–241. [accessed 2025 Oct 2]. Available from: https://doi.org/10.1007/978-3-319-24574-4_28

Schilling, M.P., Schmelzer, S., Klinger, L. & Reischl, M. (2022) KaIDA: a modular tool for assisting image annotation in deep learning. *Journal of Integrative Bioinformatics*, 19(4). Available from: https://doi.org/10.1515/jib-2022-0018

Shirali, H., Ascenzi, A., Wührl, L., Beyer, N., di Lorenzo, N., Vaccarella, E. et al. (2025a) InsectMorphoAI: a deep learning-based software for automated estimation of insect length, volume, and biomass. *bioRxiv*. Available from: https://doi.org/10.1101/2025.05.22.655251

Shirali, H., Hübner, J., Both, R., Raupach, M., Reischl, M., Schmidt, S. et al. (2024) Image-based recognition of parasitoid wasps using advanced neural networks. *Invertebrate Systematics*, 38(6), IS24011. https://www.publish.csiro.au/is/IS24011. Available from: https://doi.org/10.1071/IS24011

Shirali, H., Wührl, L., Lee, L., Klug, N., Meier, R., Pylatiuk, C. et al. (2025b) Dark-taxa-triage-dl. GitLab, Karlsruhe Institute of Technology. https://gitlab.kit.edu/kit/iai/ber/dark-taxa-triage-dl

Shirali, H., Wührl, L., Lee, L., Klug, N., Meier, R., Pylatiuk, C. et al. (2025c) Image-based Orientation, Sex, and Anatomical Segmentation of Phoridae (Diptera). Zenodo. Version 2. https://zenodo.org/records/18429938

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

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

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

Stork, N.E. (2018) How many species of insects and other terrestrial arthropods are there on earth? *Annual Review of Entomology*, 63, 31–45. Available from: https://doi.org/10.1146/annurev-ento-020117-043348

Tan, M., & Le, Q.V. (2019). EfficientNet: rethinking model scaling for convolutional neural networks. [accessed 2025 Oct 2]. https://arxiv.org/abs/1905.11946. https://doi.org/10.48550/ARXIV.1905.11946.

Toulkeridou, E., Gutierrez, C.E., Baum, D., Doya, K. & Economo, E.P. (2023) Automated segmentation of insect anatomy from micro-ct images using deep learning. *Natural Sciences*, 3(4), e20230010. Available from: https://doi.org/10.1002/ntls.20230010

Tuda, M. & Luna-Maldonado, A.I. (2020) Image-based insect species and gender classification by trained supervised machine learning algorithms. *Ecological Informatics*, 60, 101135. Available from: https://doi.org/10.1016/j.ecoinf.2020.101135

Van Dam, A.R. & Štarhová Serbina, L. (2025) Descriptron: artificial intelligence for automating taxonomic species descriptions with a user-friendly software package. *Systematic Entomology*, 51(1), e70005. Available from: https://doi.org/10.1111/syen.70005

Wägele, J.W., Bodesheim, P., Bourlat, S.J., Denzler, J., Diepenbroek, M., Fonseca, V. et al. (2022) Towards a multisensor station for automated biodiversity monitoring. *Basic and Applied Ecology*, 59, 105–138. Available from: https://doi.org/10.1016/j.baae.2022.01.003

Wührl, L., Pylatiuk, C., Giersch, M., Lapp, F., von Rintelen, T., Balke, M. et al. (2022) DiversityScanner: robotic handling of small invertebrates with machine learning methods. *Molecular Ecology Resources*, 22(4), 1626–1638. Available from: 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. Available from: https://doi.org/10.1109/ACCESS.2024.3355272

Wührl, L., Rotmann, K., Meier, R., Hartop, E., Klug, N. & Pylatiuk, C. (2023) DiversityScanner-360°: an automated system for digitizing invertebrate bulk samples. In: *2023 3rd International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*. Singapore: IEEE, pp. 226–230. [accessed 2026 Feb 17]. https://ieeexplore.ieee.org/document/10601274. Available from: https://doi.org/10.1109/RAAI59955.2023.10601274

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Data S1: Independent expert validation dataset including visibility assessment, sex classification comparisons and orientation classification comparisons for all test images (Excel file with 3 sheets).

Figure S1: Training and validation curves for the best-performing models. (a) Loss curves for the balanced YOLOv8x-cls model on the orientation classification task. (b) Loss curves for the BEiTv2 model on the sex classification task.

Figure S2: Confusion matrices for the best-performing classification models. (a) Confusion matrix for the BEiTv2 model on the sex task. (b) Confusion matrix for the balanced YOLOv8x-cls model on the orientation task.

Figure S3: Pixel count distribution for each anatomical class in the segmentation dataset showing class imbalance.

Figure S4: Training and validation curves for the best-performing segmentation model. (A) Loss and (B) Intersection over Union (IoU) curves for the U-Net with an EfficientNetB0 backbone.

How to cite this article: Shirali, H., Wührl, L., Lee, L., Klug, N., Meier, R., Pylatiuk, C. et al. (2026) Automated specimen triage for dark taxa: Deep learning enables orientation, sex identification and anatomical segmentation from robotic imaging. *Systematic Entomology*, 51(1), e70039. Available from: https://doi.org/10.1111/syen.70039