WOLO: Wilson Only Looks Once – Estimating ant body mass from reference-free images using deep convolutional neural networks

Fabian Plum[1*], Lena Plum[2], Corvin Bischoff[1], David Labonte[1*]

[1]Department of Bioengineering, Imperial College London, London, United Kingdom.

[2]Federal Highway Research Institute Germany (BASt), Bergisch Gladbach, Germany.

*Corresponding author(s). E-mail(s): fabian.plum18@imperial.ac.uk; d.labonte@imperial.ac.uk;

Abstract

Size estimation is a hard computer vision problem with widespread applications in quality control in manufacturing and processing plants, livestock management, and studies on animal behaviour. Typically, image-based size estimation is facilitated by either well-controlled imaging conditions, the provision of global cues, or both. Reference-free size estimation is challenging, because objects of vastly different sizes can appear identical if they are of similar shape. Here, we attempt to implement automated and reference-free body size estimation to facilitate large-scale experimental work in a key model species in sociobiology: the leaf-cutter ants. Leaf-cutter ants are a suitable testbed for reference-free size-estimation, because their workers differ vastly in both size and shape; in principle, it is therefore possible to infer body mass, a proxy for size, from relative body proportions alone. Inspired by earlier work by E.O. Wilson, who trained himself to discern ant worker size from visual cues alone, we used various deep learning techniques to achieve the same feat automatically, quickly, and at scale from a single reference image: *Wilson Only Looks Once* (WOLO). Utilizing over 3 million hand-annotated and computer-generated images, a set of deep neural networks—including regressors, classifiers, and detectors—were trained to estimate the body mass of ants from image cut-outs. The WOLO networks approximately matched human performance, measured for a small group of both experts and non-experts, but were about 1000 times faster. Further refinement may enable accurate, high-throughput, and non-intrusive body weight estimation at scale, and so eventually contribute to a more nuanced and comprehensive understanding of the complex division of labour that characterises polymorphic insect societies.

Introduction

Image-based size estimation is an important computer vision task, rendered challenging by the complexity and variability of visual cues. Applications range from agriculture and robotics to animal behavioural research [1–8]. Although different in motivation, these applications have in common the need to unify the appearance of the target subjects across images, and to provide pre-processed information for accurate inference [1, 3, 7, 9]. Image-based size estimation typically focuses on robotic or agricultural subjects [1–8]; popular methods include convolutional network architectures that produce intermediate pose estimates, or binary image segmentations, to provide approximate measurements from which size-estimates can be extracted [5, 6, 8]. In agricultural settings, e.g. in fruit processing plants [2, 4] or in livestock rearing [5, 6, 8], the recording environments are typically standardised, so that images have consistent camera-subject angles and camera-subject distances, which reduces task complexity. As visual information alone is often insufficient to accurately estimate size, it is common practise to include absolute scale information in the images [1, 3, 7, 9], for example in form of reference objects of known size. Radar, sonar, or infrared light, can help address the same problem, because weight can then be estimated from coarse 3D object reconstructions [10–12]. Completely reference-free size-estimation, however, is rare [6].

The key challenge with reference-free weight estimation is that visually similar objects may be of vastly different sizes; a tiny toy car can be readily confused with a real-sized car, through manipulation of image magnification [3]. As a consequence, global cues are typically vital for robust size estimation, but they cannot always be provided, and at the very least reduce application versatility. One scenario where reference-free weight estimation should at least in principle be possible is where object weight varies with object shape; size may then be inferred solely from the object itself, through assessment of relative subject proportions. In this work, we tackle one such example: the workers of leaf-cutter ant colonies [13–15].

Leaf-cutter ants (Subtribe Attina) form complex societies comprising large numbers of sterile “worker” ants, which can vary in body mass by more than two orders of magnitude [13–18]. Leaf-cutter ant colonies present a textbook example of a division of labour in eusocial insects that transcends the division into reproductive and sterile castes: morphological differentiation is coupled with size-specific task specialisation [14, 19]. The smallest individuals (minims) primarily tend to the fungus garden, the queen and partake in brood care; medium to large sized workers (medias) cut and process plant matter; and the very largest workers (majors, often referred to as soldiers) almost exclusively partake in colony defense without contributing to the foraging efforts directly [14, 15, 17, 20]. Additional complexity arises within *Atta* colonies as the variation in worker sizes is continuous rather than split into discrete sub-castes [14, 21–23] and the tasks carried out by workers of different sizes may change with the colony feeding state, age, distance to food sources, temperature, and ontogeny of individual workers [14, 16, 22, 24–27]. The resulting task choices are hypothesised to lead to an ergonomic optimum - that is worker sizes are allocated in such a way that each task is carried out to maximise the energy available to the colony [15, 28–31]. Therefore, especially energetically demanding tasks such as foraging require appropriate worker size frequency distributions in the participating animals. To give but a few examples, size frequency distributions of foraging parties appear to be adapted to the specific requirements of the available food sources [14, 32] and are affected by the food sources' geometric properties[33]; toughness and thus required cutting force [34]; and fragment surface-area and mass [35].

Unravelling the "rules" that underlie the complex organisation of leaf-cutter ant colonies has been a long-standing challenge in sociobiology, rendered difficult by the large number of involved behaviours, and the large number of individuals per colony. In the absence of better options, researchers often resort to manual extraction and weighing of individual workers, which is time consuming, error-prone, and disruptive (see e.g. [13, 14, 36]).

To minimise disruption, E.O. Wilson trained himself to estimate leaf-cutter head width by eye, aided by a physical look-up table in form of pinned workers [14]. Wilson reported that he was able to assign ant workers into one of 24 discretised size classes with an accuracy of 90 %, with the remaining ten percent placed into adjacent classes. Inspired by this work, we here aim to train deep neural networks to perform such visual size estimation rapidly, at scale, and reference-free —that is, without provision of an absolute scale. Unlike Wilson, who had rich contextual information as well as a physical look-up table and virtually unbounded perspectives for each estimate, we attempt inference of body mass as a proxy for size from only a single cropped image sample: *Wilson Only Looks Once* (WOLO).

Results and Discussion

Body weight estimation from images is a challenging task, and usually only feasible if reference lengths or other cues are provided. In *Atta* leaf-cutter ants workers, body size variation is accompanied by changes in body shape [24, 37–43]; it is thus in principle possible to learn how to estimate body mass or other size proxies without external cues, and independent of image magnification [14]. To this end, we trained a variety of deep convolutional neural networks to estimate leaf-cutter ant body mass directly from cropped or full-frame samples. Three annotated datasets were curated for training and performance benchmarking. The main dataset consists of $3 \times 5 \times 10{,}000 = 150{,}000$ full frame image samples; 10,000 frames were recorded with three cameras, each with different perspective, and with five different recording area backgrounds. Each frame contained 20 individuals that represent one of 20 body mass classes (Supplementary

Table 1, see also Fig. 7, A), resulting in a total of 3 million cropped-frame samples; the specific set of 20 individuals differed across recording backgrounds, leading to a total of 100 distinct ant workers. Computer-generated images, generated and annotated with *scAnt*[44] and *replicAnt*[45] were used to augment training datasets as discussed further below (Fig. 7), and two additional test datasets, procured to represent different inference scenarios, provide Out-Of-Distribution (OOD) test data (Fig. 1).

Network performance was evaluated in terms of prediction accuracy, precision and bias. We distinguish between qualitative accuracy, defined as the ability to rank individuals correctly by size and quantified through Spearman's Rank Correlation Coefficient (SRCC) between estimated and ground truth body mass, and quantitative accuracy, as quantified through the Mean Relative Percentage Error (MAPE). Both metrics are evaluated on mean or mode predictions per unique individual as appropriate. Precision is assessed through the coefficient of variation (CoV), this time evaluated on all predictions for the same individual. Last, network bias is evaluated through the SRCC of accuracy with body mass, zero only for unbiased networks. Detailed descriptions of dataset curation, model fitting, loss functions and performance metrics can be found in the methods. In total, 98 networks were evaluated. These networks differed in training and inference mode, training data, and regularisation. For the sake of clarity and brevity, the performance of all networks is comprehensively summarised in Supplementary Table S2, and only the main trends and key results are summarised here.

Regressors rank order well, but suffer from accuracy bias

The most natural implementation of the body weight estimation problem is arguably regression. This approach was realised in form of a cropped-frame deep regressor, based on the XceptionNet architecture [47]. The regressor had a frozen network backbone as a feature extractor, followed by two fully connected layers of 4096 nodes each, and a single output node as a head (Fig. 8). The regressor network was trained for a total of 50 EPOCHS with the Adam optimiser [48], and using the mean square error (MSE) as the loss function. Network performance was evaluated on the final 20 % of the main dataset, consisting of 500,000 samples withheld at training time. A regressor network trained on raw body masses achieved good qualitative accuracy (SRCC = 0.826), but lacked quantitative accuracy (MAPE = 132.3 %). Worse still, this relative error varied substantially and systematically with worker size: the body mass of small individuals was overestimated with large relative errors, and the body mass of large individuals was underestimated with small relative errors (Fig. 2, B). In other words, the regressor's accuracy was strongly biased (SRCC_{acc} = -0.927). One way reduce this bias is to train the regressor network on log10-transformed instead of raw body masses instead, i. e. to minimise relative instead of absolute errors (Fig. 8 E). This approach almost doubled the quantitative prediction accuracy (MAPE = 67.5 %), and achieved a higher qualitative accuracy (SRCC = 0.90) (Fig. 2 C & D). However, a strong bias in prediction accuracy remained (SRCC_{acc} = -0.9). The regressor trained on log10-transformed data produced a slightly higher CoV than the regressor trained on raw mass data, 0.424 and 0.358 respectively, indicating a decrease in precision.

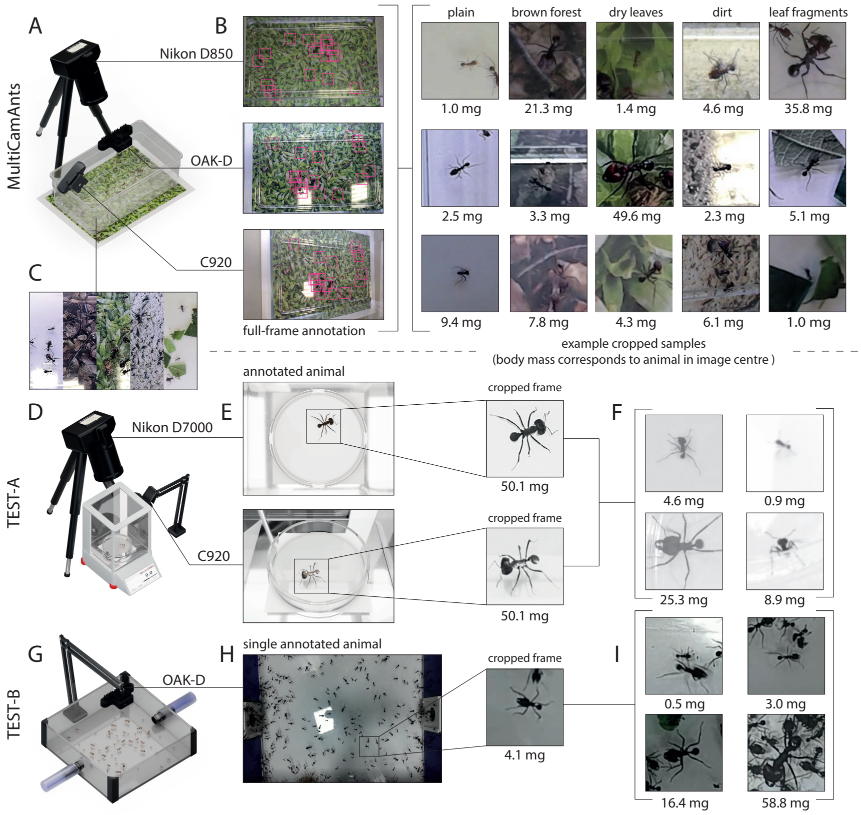


Fig. 1 Curation of training, validation, and test datasets. (A) Three synchronised cameras—a Nikon D850 with a 18-105 mm Nikkor lens, an OAK-D machine vision and a Logitech C920—were used to record synchronised videos of 20 *Atta vollenweideri* (Forel, 1893) leaf-cutter ant workers, ranging in body mass from 1 to 50 mg. (B) The cameras recorded images of the same individuals from three unique perspectives, and with different magnification (Fig. 7, and Supplementary Table S1 for exact worker weights). 10,000 frames were annotated semi-automatically using *OmniTrax* [46]. To streamline data processing, only the top-down OAK-D recordings were annotated, and a custom-written perspective conversion script was used to translate the tracks, indicated by red bounding boxes, into the other two camera views. (C) Further variation was introduced by means of an interchangeable background: either a default plain background, a textured brown forest floor, dry leaves, dirt, or a plain background with cluttered leaf-fragments. The 20 individuals always covered the same weight range, but each background had a unique set of individuals. (B) The resulting dataset contained $3 \times 5 \times 10{,}000 = 150{,}000$ labelled frames, each containing 20 individuals, resulting in a total of 3,000,000 cropped image samples (examples on the right). The first 80% of the full frame and cropped datasets were used as training data; the remaining 20% served as unseen validation data. Two out-of-distribution datasets were curated for benchmarking. (D) Dataset *Test A* was recorded with a Nikon D7000 camera, equipped with a micro Nikkor 105 mm lens, oriented top-down, and a Logitech C920 camera at an angle of approximately 30% to the vertical. Single leaf-cutter ant workers were put into a Petri dish placed on an ultra fine scale, and (E) between 20 to 50 monochromatic frames were captured for each individual resulting in (F) 4,944 cropped and annotated samples. (G) *Test-B* was recorded with an OAK-D camera positioned above a crowded container that served as a section of a laboratory foraging trail. (H) Individual workers were annotated with the manual tracking module of *OmniTrax* [46]; (I) a total of 30,526 cropped RGB samples were extracted.

It is not immediately obvious why regressor network accuracy was strongly biased. One possible explanation is that networks by definition never see training samples with a body mass outside the target domain, so that the mass of small workers is substantially more likely to be over- than underestimated; mass-estimation in large workers suffers from the opposite problem, and this argument is consistent with the data (Fig. 2). The strong bias may also simply reflect that the mass-estimation problem is hard, so that networks may be unable to predict masses with consistently high confidence. Performance may then be optimised if estimates regress towards the sample mean, which prevents error inflation due to outliers. In support of this argument, regressors trained on raw data were most accurate around the sample mean body mass of 13 mg, whereas the regressors trained on log10-transformed data were most accurate around the log10-transformed sample mean body mass of about 7 mg (Fig. 2). Motivated by this interpretation, we next implemented the weight estimation problem as a classification task: even if classifiers badly miss-classified workers from time to time, as long as they classify them correctly *most of the time*, accuracy bias should be reduced.

Classifiers are more accurate and unbiased, but produce larger fluctuations in their predictions

To implement a body mass classifier, we took inspiration from E.O. Wilson, who taught himself to classify *Atta sexdens* (Linnaeus, 1758) workers into one of 24 head width classes by eye [14]. Instead of using head width as a proxy for size on a linear scale, we defined 20 body mass classes, spanning the weight range [1, 50] mg, with class-centres equi-distant in log10-space (Supplementary Table S1). Classifiers also operated on cropped frame samples, and their architecture was identical to that of regressors, apart from the head, which was replaced by one node per target class (see fig. 8). Classifier networks were trained using cross entropy loss and one-hot encoded discretised class labels; their accuracy was evaluated on the prediction mode per unique individual—the appropriate metric of central tendency for categorical data.

Classifiers achieved MAPEs as low as 33 %. Recognising that mass-discretisation carries an inherent minimal error of about 6 % (due to the within-class body mass distribution, see methods), this corresponds to an improvement of factors between two and three compared to regressors. Remarkably, classifier accuracy was also independent of body mass ($\mathrm{SRCC}_{\mathrm{acc}}$ = -0.098). However, the price paid for these improvements was a reduction in prediction precision: the coefficient of variation, computed across predictions for the same individual across frames, almost doubled from 0.42 for the regressor trained on log10-transformed data to 0.80, most clearly visible in discretised confusion matrices (Fig. 3). As a more appropriate metric to quantify classifier prediction stability, this fluctuation corresponds to 47.5% of classifications being assigned to their respective individual prediction mode (see Methods as well as Supplementary Table S2 for details). This effect is consistent with the suggestion that regressor accuracy is poor to avoid influential outliers.

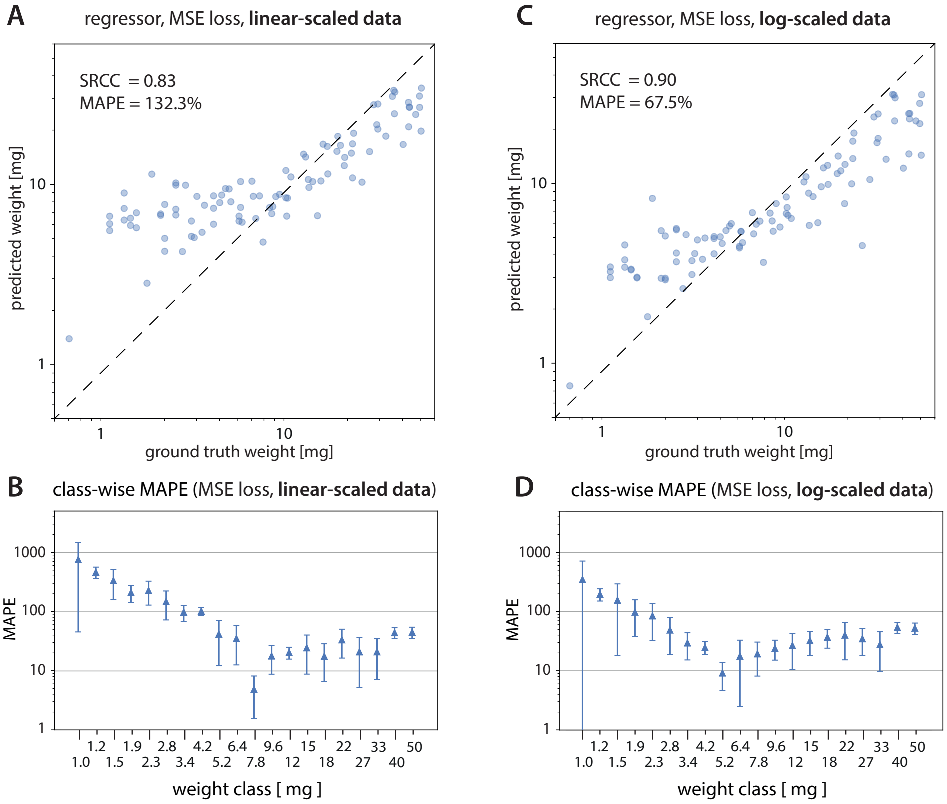


Fig. 2 Deep regressor networks, trained on 2.5 million cropped frame samples of leaf-cutter ant workers of different body mass, can rank individuals by their mass with reasonably high qualitative accuracy, as assessed by Spearman's Rank Correlation Coefficient (SRCC). However, prediction accuracy is biased, and quantitative accuracy is poor for both small and large individuals, as quantified by the Mean Average Percentage Error (MAPE). All results are from withheld within-domain data. (A) & (C) show parity plots of the mean body mass prediction vs ground-truth for regressors trained on raw (untransformed) and log10-transformed body masses, respectively. The dashed line shows the parity line, which indicates perfect prediction accuracy. (B) & (D) show bar plot of the class-wise MAPE, grouped in bins with equal centre-to-centre distance in log10-space for the same regressors. Log10-transformation of body masses improved the qualitative and quantitative accuracy, and reduced the bias, which however remained significant. Regressors generally achieved the best accuracy for individuals with a body mass close to the sample mean (13 vs 7 mg, for raw and log10-transformed body mass, respectively). This observation indicates that accuracy bias may be caused by a "regression-to-the-mean", incentivised by the need to avoid prediction outliers, which may arise from low prediction confidence.

Regressors have poor accuracy and are strongly biased, and classifiers have high accuracy and are unbiased, but suffer from outliers. These differences likely arise from the different statistical elements that determine performance across the two approaches, which rely on ordinal vs categorial information. In an attempt to combine the best of both worlds, we introduced ordinal characteristics into classification through class-relationship aware label smoothing (CRALS). The main idea is to replace one-hot linear encoding with a Gaussian activation profile, so that miss-classification in adjacent classes is penalised less than miss-classification in classes further away. The strength of this effect can be tuned through the parameter σ, the standard deviation of the Gaussian distribution (Fig. 8 D). Effectively, CRALS acts as a regularisation technique: for very small σ, the approach resembles pure classification, and for very large σ, it will approach discretised regression.

In support of this argument, networks trained with CRALS typically achieved better qualitative and quantitative accuracy than regressors, and less prone to producing distant outliers than pure classifiers (Fig. 3). Label smoothing rendered predictions "fuzzier" (Fig. 3 F), in the sense that they clustered more closely around the mode, as observed for regressors (Fig. 3 B). However, and in contrast to regressors, the mode remained unbiased, and tightly clustered around the ground truth, as observed for classifiers (Fig. 3 D). Another advantage of CRALS as a regularisation technique, while slightly decreasing categorical classification accuracy on validation examples, lies in a better retention of performance across metrics in OOD cases. The strength of these effects generally varied with the magnitude of σ (see Supplementary Table S1 for full details). Classifiers performed well on within-distribution data: the best network had a MAPE of about 30.9 %, independent of size. However, the ideal network also has to be able to generalise; that is, it ought to be robust to variations in recording settings, and work on OOD data without a strong drop in performance. To test this ability, two OOD datasets were procured. Test A consists of images of single workers on a white background, without any other objects in-frame (Fig. 1 D–F); Test B comprises cluttered images of busy laboratory foraging trails, including individual worker overlap and partial occlusion (Fig. 1 G-I). Even the best classifier performed poorly on Test A; the prediction error almost doubled, and, perhaps surprisingly, a strong accuracy bias returned (Fig. 4). The qualitative accuracy however remained high (SRCC = 0.87). The performance on Test B was much better, and had an error comparable to the within-distribution performance (MAPE = 44.2 %). However, the precision suffered significantly, i. e. the predictions were much noiser; we thus conclude that prediction robustness and network generalisability were somewhat wanting, likely indicating overfitting. Other potential sources of error arise from different colour spaces in the recorded datasets and motion blur (TEST A), as well as much higher degrees of individual overlap and occlusion (TEST B)

Synthetic data increases robustness in terms of qualitative accuracy

Prediction robustness requires that networks learn generalisable cues (see e.g. [49–52]), and this ability can be fostered through extensive augmentation or large training datasets with maximal image variability. To implement augmentation on a large scale, we produced thousands of computer-generate images that possesses far greater variability in appearance than the original training data. Synthetic datasets were generated with *replicAnt*, a computational data generation pipeline implemented in

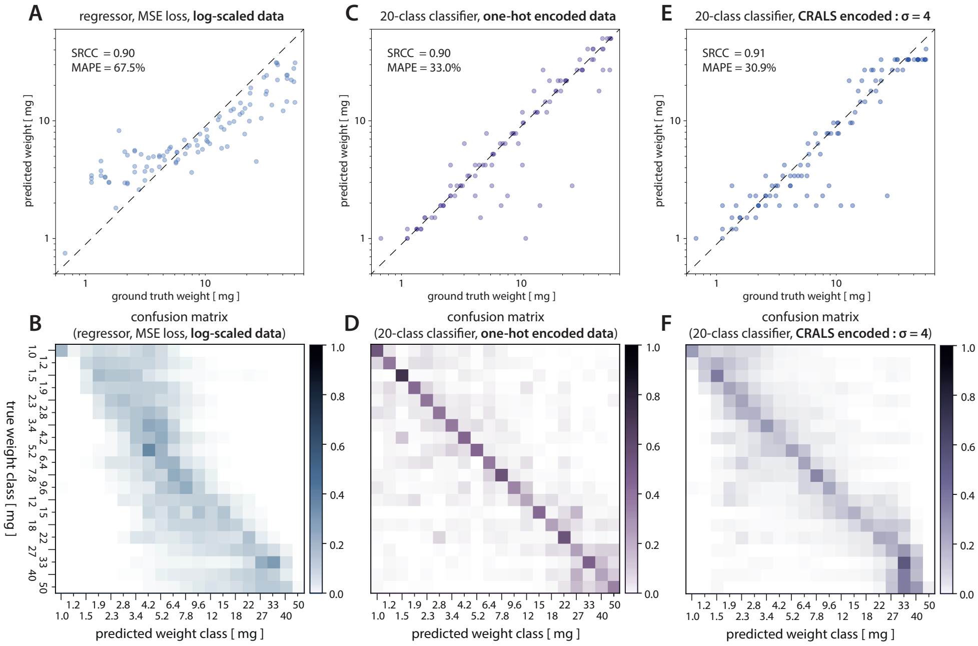


Fig. 3 Classifiers achieved high accuracy, have no prediction bias, but are more prone to outliers. The top row (A, C, E) shows parity plots for within-distribution validation data of (A) the best regressor, trained on log10-transformed weight data; (C) a classifier, trained on 20 class discretised data with one-hot encoded labels; and (E) a classifier, trained on discretised data with class-relationship aware Gaussian label smoothing ($\sigma = 4$). Datapoints represent the arithmetic mean for the regressor, but the mode for the classifiers, evaluated across all frames for each unique individual. The bottom row (B, D, F) shows the corresponding confusion matrices. Note that for the regressor, predictions cluster more tightly around the class with highest activation, but this class itself is biased with respect to the ground truth. Classifiers, in turn, stray further from the class mode on occasion, but show practically no prediction bias. Label smoothing introduces ordinal characteristics to classification, and so effectively acts as a regularisation technique that can control the trade-off between these two effects.

Unreal Engine 5 and Python [45]. *replicAnt* takes textured and rigged 3D animal models as an input, and places simulated populations of these models into complex, procedurally generated environments. From these environments, computer-annotated images can be exported, which then serve as training data supplement. Two synthetic datasets, containing a total of 200,000 full-frame annotated images from digital populations of 200 simulated individuals were created, providing a further 1,630,000 cropped frame image samples (see methods and Fig. 7).

Networks trained exclusively with synthetic data had poor quantitative accuracy, but did well in rank-ordering, suggesting that compression artefacts or absolute frame occupancy, and not shape difference, are the strongest indicators for individual size, at least for the resolution and quality of images tested here. This observation is consistent with the generally poor performance on the seemingly simple Test A dataset, which contains minimal cues other than the individual itself (see also below). In line with expectation, supplementing the training data with synthetic images moderately improved OOD performance, and most notably increased prediction precision (Fig. 4, and Supplementary Fig. S10 for a more comprehensive evaluation). Images from Test A turned out to be particularly challenging, likely because they provide no contextual information. For the same reason, they are however also somewhat artificial constructs; few, if any, real use case will resemble these imaging conditions. Most realistic applications will instead involve images that contain multiple individuals alongside other objects such as leaf litter or dirt. It will often be possible to keep camera magnification and orientation fixed, so that reliable contextual information independent of the focal individual will almost always be available. To investigate whether networks can learn to use contextual information, we next implemented a full-frame detector that localises and classifies individuals in a single pass, and thus has access to more rich information at training time.

Contextual information improves performance at all scales

To explore the ability of networks to learn from global reference cues, a full-frame custom YOLOv4 detector was given access to all contextual information of a full-frame [53]; temporal information on adjacent frames was withheld. Detector models were trained on full-frame datasets, both real and synthetic, and their performance was quantified on both the full-frame validation dataset, and the two cropped-frame OOD datasets (see methods). Detectors produced the highest achieved categorical accuracies, as 20 and as 5 class variants (see Appendix 2). Overall, however, detectors performed comparable to their equivalent classifier networks on validation data, in terms of MAPE and SRCC, and displayed no obvious prediction bias. However, detector performance drastically deteriorated when contextual information was absent, i. e. on OOD data (Fig. 5). It is not trivial to decide whether this performance drop reflects the fact that the network learned to improve inference performance through the use of contextual information, absent or reduced in the OOD data, or whether it simply indicates overfitting to training data. Without doubt, any network will always benefit from additional refinement with annotated samples from the target domain. However, irrespective of overfitting, which may well affect all networks, a meaningful advantage of a one-pass detector is a substantial reduction in inference time (see discussion below).

The best networks outperform weakly trained humans

To put network performance into perspective, we conducted a small pilot study in which 14 human participants were asked to estimate body mass from images. Both leaf-cutter ant experts (those who self-declared to work regularly with leaf-cutter ants), and non-experts were included. To reduce task complexity, the body weight estimation task was implemented in form of a 5-class classification problem, again with classes that ranged from [1,50] mg, and class centres that were equi-distant in log10-space. To compare human and network performance more directly, we trained 5-class classifier networks, identical to the original 20 class classifier, bar the adjusted classifier

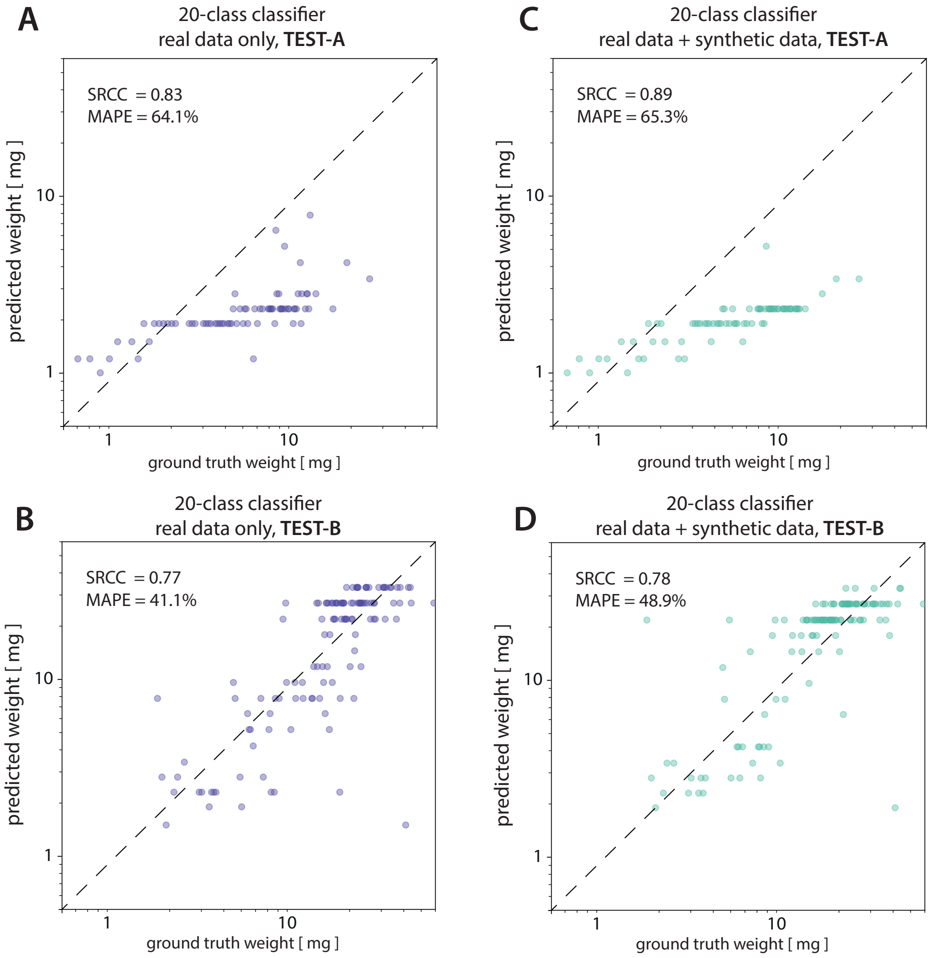


Fig. 4 Synthetic data moderately increases the ability of networks to retain qualitative performance in Out-Of-Distribution (OOD) data. Performance comparison of a 20-class classifier trained with Class-Relationship Aware Label Smoothing (CRALS) on OOD data. (A & B) show parity plots of the model trained on real data only. (C & D) show the parity plots of the same model, augmented with synthetic samples from the combined "synth standard" and "synth simple" dataset. Data points represent the prediction mode per unique individual.

head. Human participants generally performed similar or slightly worse than the best implemented networks in terms of qualitative and quantitative accuracy; experts performed consistently better than non-experts (Fig. 6 A-C). In light of the small sample size, neither human bias nor accuracy can be reliably estimated. Wilson reported an accuracy of 90 % on a classification task with 24 classes, almost seven times higher than the best human performance in this study, relative to a random guess [14] ([see also 45]). It is unclear to what extent this difference in performance stems from sheer

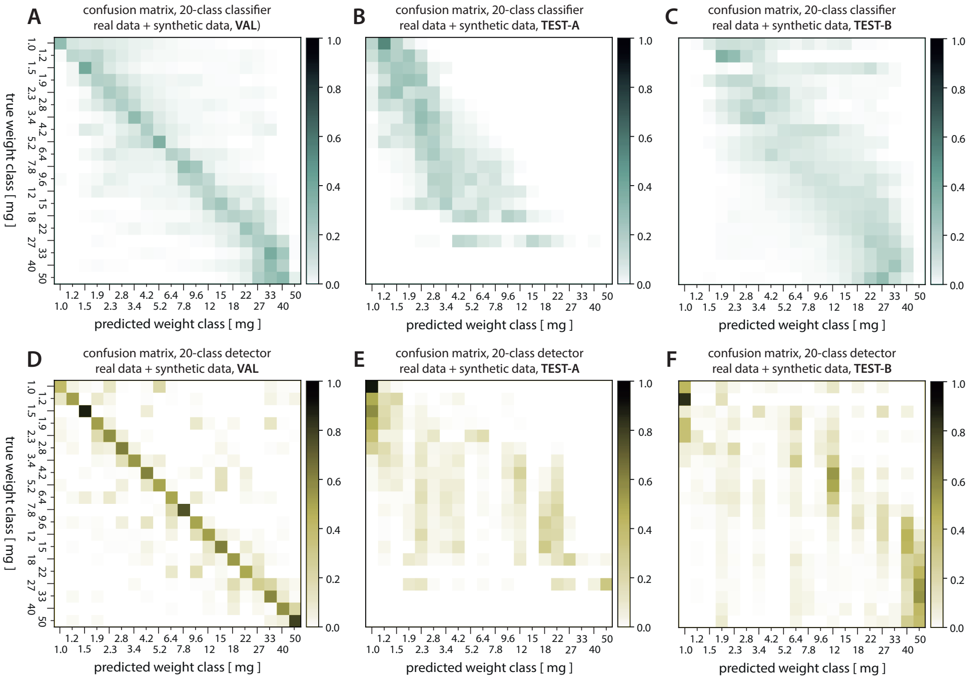


Fig. 5 Full-frame detection and weight estimation in principle provides access to contextual information, and may thus improve prediction performance. However, estimation performance of the detector deteriorates drastically on Out-Of-Distribution (OOD) data, indicating that the strong performance may be the result of overfitting. The top row shows confusion matrices for the best 20 class classifier (label smoothing applied, $\sigma = 4$, trained on mixed MultiCamAnts and all synthetic data) for (A) the within-distribution validation split of the MultiCamAnts dataset; (B) Test A; and (C) Test B OOD datasets, respectively. Both the qualitative and quantitative accuracy are worse on OOD data (MAPE in VAL = 36.35 %, Test A = 65.3 %, Test B = 49.0 %; SRCC in VAL = 0.915, Test A = 0.886, Test B = 0.775). The bottom row shows confusion matrices for a 20 class detector trained on equivalent full-frame mixed real and all synthetic data; (D) shows the within-distribution validation split of the MultiCamAnts dataset; (E) and (F) are for Test A and B, respectively (MAPE in VAL = 45.5 %, Test A = 93.3 %, Test B = 133.9 %); SRCC in VAL = 0.916, Test A = 0.670, Test B = 0.730).

training, innate skill, or from richer contextual information: Wilson observed foraging workers in their natural environment for prolonged periods, and had an extensive physical look-up table at hand. In contrast, human participants only received a single low-resolution cropped image. We thus stress that we do not claim that the pilot study provides a reasonable indication of the upper limit of human performance; we do however believe that it supports both the weaker conclusion that the task itself is hard, and that the networks are able to learn a considerable amount from the training data provided.

In an effort to understand the limits to both human and network performance further, it is instructive to inspect image samples that were frequently classified correctly or

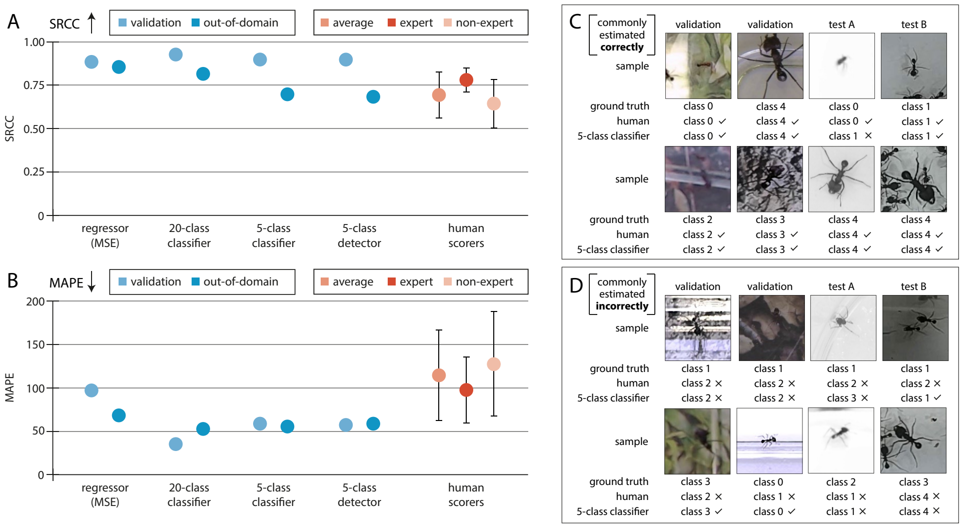


Fig. 6 (A) The best networks achieved similar or better performance than humans on the unseen validation data, and on Test A & B. (B) The best networks consistently achieved lower MAPEs in Test data. (C) A collection of image samples that were typically classified correctly by human annotators, compared with the classification of the best performing 5-class classifier. (D) A collection of image samples that were frequently classified incorrectly by human annotators, compared with the classification of the best performing 5-class classifier. Note that the computational models have seen substantially more training samples; combined with the small number of human participants ($n = 14$), this pilot study only provides a rough estimate of human performance that indicates the difficulty of reference-free weight estimation.

incorrectly, respectively (Fig. 6 D-E). Human participants and network variants struggled with similar images: workers that were largely out-of-focus, the general presence of image noise, unusual poses, or deviations from top-down perspectives that distort or obscure morphological landmarks such as head widths or leg lengths all rendered the inference problem harder. A notable exception to this rule appear to be images of the very smallest ant workers, on which networks performed poorly, but humans often did well, likely because they (correctly) inferred that the low magnification and high noise imply a small animal size (Fig. 6 D, top-row, Test A, class 0). Although the same may be expected for networks, this is exactly one of the principal advantages of synthetic data, because it can bypass such generally undesired biases. Irrespective of prediction quality, the main difference between human and computational classification lies elsewhere: Human annotators took on average six seconds to classify an image sample, whereas the trained models can perform about 1000 times as many predictions in the same time; they thus vastly outperform humans in terms of speed.

Conclusion and Outlook

Inspired by the work of E.O. Wilson, who trained himself to estimate leaf-cutter ant worker body mass by eye [14], the aim of this work was to investigate the possibility to infer ant body mass from reference-free images using deep convolutional neural networks. Because size-differences in leaf-cutter ants are associated with differences in shape, this aim is achievable in principle.

Despite relatively large amounts of training data, and the exploration of diverse inference approaches, the performance of even the best networks remained below the self-reported accuracy of E.O. Wilson; it was, however, comparable to that of human annotators that have had less dedicated and extensive training. Irrespective of these differences, the best networks had a size-independent relative error of about 30 %, which may well be good enough for many purposes. The key advantage of the automated approaches presented here is their vastly superior speed; any loss in accuracy can likely be balanced by an increase in sample size, so that statistical power remains sufficient. Indeed, given that *Atta* trails readily contain thousands of individuals, automated mass-estimation may well be the only realistic and affordable option to obtain data on size-frequency distributions at the required scale. Where higher accuracy is needed, the most effective route may be the provision of an absolute scale, as done in previous related work [2, 3, 7]. Worker size can then be measured directly from the images, e. g., through body length or pixel number, with either pose estimators [54–56] or binary masks [2, 4, 5], respectively. However, for all its advantages, this approach is not without problems: parallax errors can only be managed through tight control of recording conditions, so reducing flexibility; and key body markers may often be occluded, be it by leaf-fragments carried, or by other individuals that cross paths on busy foraging trails. Notwithstanding these difficulties, deep learning-based reference-free weight estimation has the potential to become a valuable asset in behavioural research on leaf-cutter ants and other polymorphic insects. Ample opportunity for algorithmic improvement exists, and the presented inference approaches present a promising step towards more nuanced, efficient, and non-intrusive methods in the study of fascinating social organisation of complex insect societies.

Apart from its potential practical application in leaf-cutter ant research, our work yielded insights of relevance for reference-free weight estimation more broadly. First, and perhaps unsurprisingly, contextual information appeared to improve prediction performance. As a result, detectors that work on images that contain multiple individuals appeared to generally outperform networks that infer mass from image cut-outs. Second, class-aware label-smoothing as a regularisation technique can help reduce over-fitting during training time, and increase correlation between different performance metrics. Third, the addition of synthetic data can improve network robustness, i. e. inform the training of networks which generalise better to unseen scenarios [see also 45], here, specifically regarding a network's ability to rank-order individuals by size but less in absolute accuracy terms. Fourth, the choice of performance metric is not trivial, and standard parameters such as the MAPE can introduce significant biases. Others have therefore recommended to select models based on their R^2 [57]; future work will have to carefully and systematically address the problem of performance metric and model selection in effectively ordinal classification tasks.

Methods

This study aims to use deep learning-based computer vision approaches for automated body weight estimation of leaf-cutter ant workers. To this end, training and benchmark datasets were curated, inference approaches implemented, and their performance evaluated. In addition, a small study with human participants was conducted to provide an indication of baseline performance. These essential methodological aspects are described in detail below.

Datasets

Training and benchmark datasets

Three datasets were curated (Fig. 1): (1) a complex multi-animal dataset, recorded with three different synchronised cameras, from three different perspectives, on five different backgrounds, and with varying degrees of leaf clutter (MultiCamAnts, Fig. 1 A-C); (2) a simpler single-animal test dataset, recorded with two different cameras, from two different perspectives, and on neutral background (Test-A, Fig. 1 D-F); and (3) a top-down multi-animal dataset, recorded with a machine-vision camera, oriented top-down above a busy ant foraging trail (Fig. 1 G-I). All cropped-frame training and benchmark datasets are available via Zenodo (https://zenodo.org/records/11167521). For full-frame datasets, please contact the corresponding author.

(**1**) The MultiCamAnts recordings served as the primary dataset. Three cameras—a Nikon D850 with a Nikkor 18-105 mm lens, a OAK-D machine vision camera, and a Logitech C920—were used to record images of ants that moved inside an acrylic container that served as recording arena (250 mm x 150 mm x 90 mm). Videos were time-synchronised by triggering a Nikon SB-700 AF Speedlight Flash Unit, once before animals were placed into the arena, and then again after 10,000 frames had been captured; these two time points were used to synchronise the videos using After effects (CC version 2023, Adobe Inc.). The visual appearance was varied by exchanging the arena background, and by scattering leaf-fragments, such as to emulate the appearance of foraging trails (see Fig 1 C).

Five sets of 20 ant workers were taken from the foraging containers of a mature laboratory colony of *Atta vollenweideri* (Forel 1893) leaf-cutter ants, housed in a climate chamber at $25°C$ and 60% humidity; individuals were weighed with a precision scale (OBX-223 Ohaus Explorer Precision Balance, ± 0.1 mg resolution), and sampling continued until representative specimens for each of 20 body mass "classes" had been collected; class centres were chosen such that they were approximately equidistant in log10-space, and covered the weight range $[1, 50]$ mg (see Fig. 7 B). 20 ant workers at a time were then placed into the recording arena for each background, and in order of ascending body mass; subsequent identification was thus possible without application of physical markers.

A total of 10,000 frames were captured for each of five recordings; one for each background, and each with a different set of 20 individuals. These frames were subsequently annotated with *OmniTrax*, a deep learning-driven multi-animal tracking add-on for Blender [46]. Using user-guided semi-automatic tracking, all top-down recordings from the Oak-D camera were annotated. Subsequently, using a custom python script, the camera projections of the remaining views were solved, and the extracted homography was used to translate top-down tracks into the adjacent video perspectives. A total of 150,000 samples were labelled, exported and converted into the format required by the respective inference method (section 2). weight estimation via detections takes full frames as input, which was facilitated with custom data parsers. Classification and regression, in turn, operate on cropped frames that contain only the focal individual. Cropped frames with customisable aspect ratio and resolution were exported from *OmniTrax* via dedicated functionality; class and identity information were encoded in the filename. For full frame samples used to train detectors, a YOLO-formatted text file was generated for each frame containing the location, bounding box dimensions, and class of all visible animals. $3\,times\,5 \times 10{,}000$=150,000 samples for each of 20 individuals lead to a total of 3,000,000 cropped images. The first 80% (2,500,000) of these images were used as training-, and the final 20% (500,000) images as validation data. Splits were fixed to avoid inflation of validation scores, as can occur when training and test sets contain time-adjacent frames that are visually similar.

(**2**) In order to curate dataset *Test-A*, a camera rig was built around an ultra-fine scale(OBX-223 Ohaus Explorer Precision Balance, ± 0.1 mg precision. Fig. 1 D-E). 131 ants were chosen at random from the colony feeding box, leading to a weight distribution that roughly resembles that of natural foraging parties, ranging from 0.5 to 25 mg. Individuals were placed, one at a time, into a Petri dish with a white background, centred on the scale. They were then filmed with a Nikon D7000 DSLR, equipped with a micro Nikkor 105 mm lens facing downward, and a Logitech C920, fastened to a custom-built mount and oriented with a 30° angle relative to the vertical (Fig. 1 D). Images were captured from both cameras, using a custom python script, leveraging OpenCV [58] and libgphoto2; scale readings were recorded manually. Each camera captured 20 images per individual, with a low sampling frequency of 0.33 Hz, chosen to increase the postural variation across images of the same individual. The resulting dataset contained 4,944 cropped monochromatic image samples; about 300 images were discarded because individuals were entirely out of focus, or had unrulily escaped the recording set-up.

(**3**) A *Test-B* dataset (see 1 G-H) was curated to obtain crowded images, resembling the conditions on a busy foraging trail. An OAK-D machine vision camera was positioned above a custom-built acrylic container (280 mm x 280 mm x 90 mm), connected in-between the laboratory colony and a foraging box via a system of flexible PVC tubes (diameter 2 cm). Footage was recorded with a frame rate of 30 fps, and for a period of 20 minutes, during which the colony was actively foraging on bramble leaves provided in the foraging box. Because leaf-cutter ant workers were allowed to enter and exit the container throughout the recording period *ad libitum*, body masses needed to be determined by manual worker extraction, and subsequent weighing with the Ohaus precision scale. One at a time, 154 individuals were weighed in this way, and semi-automatically tracked in-post for ~ 200 frames, using *OmniTrax* [46]. A total of 30,526 cropped RGB frame patches were extracted using *OmniTrax*.

Synthetic datasets

A large and varied synthetic dataset, consisting of computer-generated images, was produced to augment the training split of MultiCamAnts, in the hope to increase network robustness on Out-Of-Distribution data [45]. Synthetic data were generated with *replicAnt*, a computational pipeline implemented in Unreal Engine 5 and Python [45]. *replicAnt* takes textured and rigged 3D models as an input, and places simulated populations of these models into complex, procedurally generated environments. From these environments, computer-annotated images can be exported, which can then be used as training data for machine-learning based computer vision applications, including classification, detection, tracking, 2D and 3D pose-estimation, and semantic segmentation.

To provide the required 3D input models, 20 worker ants, distinct from those used in the MultiCamAnts recordings 2) but with comparable size range, were sampled from the laboratory colony (see Supplementary Table S1). Specimens were sacrificed via freezing to produce "digital twins" with the open-source photogrammetry platform *scAnt* [44]. Specimens were prepared such that they were biting down on either a needle or thin PLA filament, so that their mandibles did not touch or overlap; this facilitated separate movement of the mandibles at a later stage (see below). Specimens were pinned in an upright position akin to their natural stance, and left to dry at room temperature for at least one week prior to scanning. This drying step ensured that the joints had sufficiently stiffened to prevent movement during scanning. Specimens were scanned with the *scAnt* hardware configuration described in Plum and Labonte 2021, the code version from the May 2023 (*dev* branch, and the default masking parameters of an improved stacking routine (https://github.com/PetteriAimonen/focus-stack). Specimens lighter than 4 mg were digitised using a 75 mm MPZ Computar lens, and a custom-built focus extension tube (see [44] for details); all other specimens were scanned using a 35 mm MPZ computar lens and a 5 mm C-mount extension ring. All processed models are available via Zenodo (https://zenodo.org/records/11167946)

All scans were performed with a colour-coded $5 \times 5 \times 5\,\text{mm}$ cube in view to enable colour calibration, and accurate re-scaling of the resulting 3D models. Scans were photogrammetrically reconstructed with 3DF Zephyr lite (v2023.03) at the highest level of detail, and with photo-consistency meshing enabled to retain fine structural details. It is not trivial to quantify photogrammetric reconstruction accuracy. As an approximate guide, *scAnt* can resolve step-changes in height of about 100 µm with an error of around 10%, and is better than 5% for steps of 500 µm [44].

Reconstructed textured meshes were exported as FBX files, and subsequently imported into Blender 3.2, to complete basic mesh cleaning (see [44]), and to apply a standardised armature (see [45]; Fig. 7 D). The rigged mesh was retopologised to decrease the number of vertices from $> 100,000$ to $\sim 10,000$, substantially reducing the subsequent computational load. The rigged and retopologised models were then scaled to their original size, using the colour-coded cube as reference, and the appearance of image textures was unified using histogram equalisation (fig 7 C). All models were then brought into *replicAnt* using the *send2Unreal* plug-in.

Within *replicAnt*, two large synthetic datasets were produced, referred to in the following as ”synth-standard” and ”synth-simple”, respectively. For both datasets, synthetic populations of 200 individuals were spawned; 10 from each original model. Within each of the 20 size classes, a randomised scale variation of 10% was applied, so that adjacent weight classes did not overlap in absolute scale. Synth-standard used the ”plants” environment provided with *replicAnt* to facilitate a complex scene generation with plant asset scatterers [45]. The resulting image samples were highly cluttered, and often included individuals spawned across multiple layers with respect to the camera plane. Synth-simple, on the other hand, used the default generation environment within *replicAnt* 1.0, with 70% of the asset scatterers removed to produce simple, dominantly planar scenes. The resulting samples were thus closer in appearance to the inference cases, albeit with greater variation in background materials and levels of occlusion. 100,000 image samples were generated for each dataset. *replicAnt*'s multi-class YOLO parser was used to export full-frame samples, and a custom-written second parser produced 128×128 px cut-out samples for every animal in every synthetically generated frame. These cut-outs were re-scaled when animals occupied a larger area to ensure that the entire animal was visible, and the subject class was encoded in the filenames. Individuals that occupied small fractions of the cut-out were centred, and basic up-sampling was applied such that the larger side of the bounding-box corresponded to at least 10% of either the width or height of the cropped image. A simple conversion script automatically sorted samples into discrete size folders, using the class information provided in the file name, and so produced the file-structure required by *TensorFlow.dataset* (see below).

All custom tools used in the curation and generation of real and synthetic data are open source, and accessible on GitHub (https://github.com/FabianPlum/WOLO).

Inference approaches

Three inference approaches were explored: image patch regression, image patch classification, and simultaneous localisation and classification on full frames (detection). All combinations of real and synthetic training samples were tested for each inference type (see Supplementary Table S2). Regression is arguably the most natural implementation of the mass-estimation problem, but suffered from prediction bias (see results). Classification appeared less prone to such bias, but comes at the cost of a minimal error defined by the difference between ground truth sample vs class-centre mass. Both regression and classification work on cropped-frame samples, returned by a separate localiser at inference time, and thus require intermediate processing steps that slow

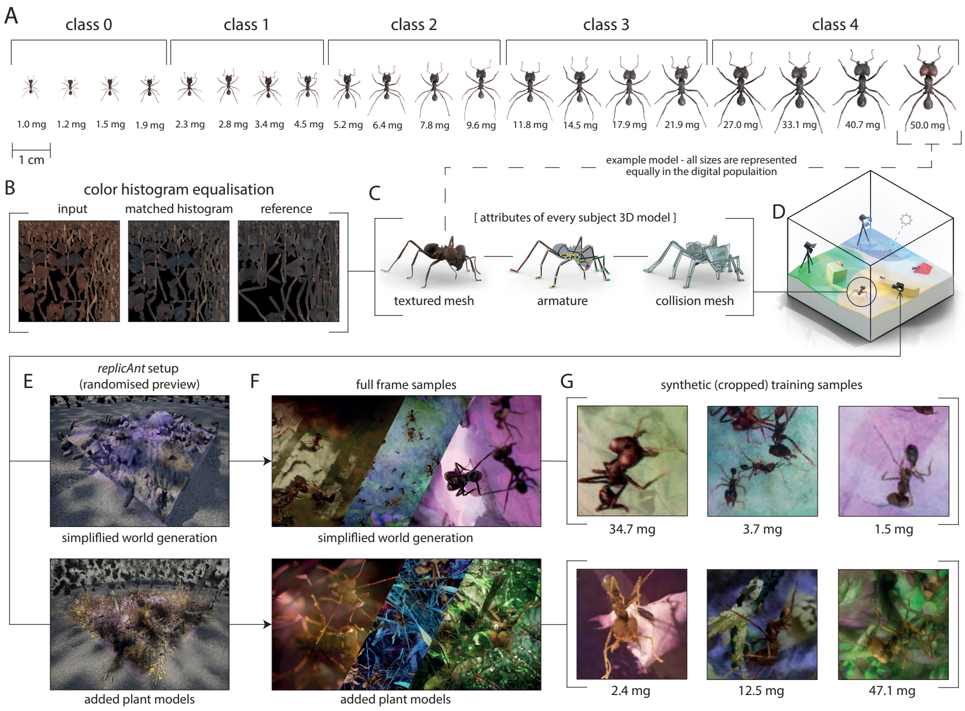


Fig. 7 To increase network robustness and generalisability, real datasets were augmented with two synthetic datasets, computer-generated with *replicAnt* [45]. (A) 20 "digital twins" of leaf-cutter ant workers across the mass-range [1, 50] mg were created with *scAnt* [44], an open-source photogrammetry platform (see Supplementary Table 1). The 20 workers were selected to approximately match the 20 class-centre body masses chosen for the classifier. For 20-class weight classification, each individual model represented one class; for 5-classification, four individuals were grouped to form one class. (B) To remove colour variation caused by subtle differences in imaging conditions, colour-histogram equalisation was performed for all texture maps, using a custom-script written in python. (C) All ant models were then retopologised and rigged in Blender (3.2), and ported to (D) *replicAnt* in Unreal engine, where a low polygonal collision mesh was computed, and the sample randomisation procedure configured. A digital population, comprising 200 individuals that differed in scale, hue, contrast, and saturation, formed the basis for (E) two synthetically generated datasets: a simplified (top) and a standard (bottom) pipeline that spawned plant assets. (F) 100,000 procedurally generated and automatically annotated full-frame examples were exported for each dataset. (G) Cropped training samples were extracted from the original full-frame samples using a custom-written parser.

down inference. Detection enables simultaneous localisation and classification of multiple individuals in a full-frame and thus avoids this problem; the network really only looks once. Full-frame detection and classification in a single pass is not only significantly faster, it also provides a network with the ability to learn from contextual scene information; the downside is the increased demand for GPU RAM (VRAM) when running inference on footage with high resolution and small individual occupancy.

Regression

Regression was performed on $128 \times 128 \times 3$ image samples (resolution in x, resolution in y, colour channels), cropped such that the thorax of the target animal was located in the image centre (Fig. 8 A). A headless Xception Net with frozen weights, pre-trained on the ImageNet (v2017) dataset, served as a feature extractor [47]. Its outputs were fed into two fully connected layers with 4096 nodes each, using ReLU as the activation function (see Fig 8 B). The final output came from a single node, which was normalised to a range between [0, 1] at training time, and then re-mapped to the original mass range to extract network predictions (see Supplementary Table S1).

The Mean Squared Error (MSE) was set as the loss function, i. e. networks were trained to minimise the absolute prediction error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

Here, n is the sample number, y_i is the prediction, and $\hat{y}_i$ the ground truth. A reasonable alternative is to minimise the relative error, which was realised by additionally training networks with log10-transformed body masses (Fig. 8 D).

Classification

Classification was performed as regression, with the sole difference that the network output is now formed by a classifier head with SoftMax activation instead of a single node. Two classifiers were trained, one with 20 classes and one with 5. 20 classes were chosen to roughly match the discretisation used by Wilson [14]; and 5 classes provide a point of comparison to human performance on the training data used in this study (see below). In both cases, classes span the weight range [1, 50] mg, and class centres were chosen such that class-centre mass were approximately equi-distant in log10-space. Both classifiers were implemented in Tensorflow (v2.9.1), and trained for 50 EPOCHS with the Adam optimiser [48], using one-hot cross-entropy as the loss function. A key difference between classification and regression is that all classification errors are equal. To render categorical classification more similar to ordinal regression, we implemented a simple class relationship-aware Gaussian label smoothing algorithm. Unlike default one-hot encoding, the label smoothing method lifts the activation of adjacent classes to the target class μ, according to a normalised Gaussian distribution with standard deviation σ (see 8 D). The normalised activation $y(x)$ of each output node is defined as:

$$y(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\frac{1}{\sum_{t=1}^{n} y(t)} \tag{2}$$

Label smoothing thus penalises incorrect predictions into a target class with a class-centre mass close to the ground truth less than an incorrect prediction into a class far away from it; classification is, in a sense, rendered more similar to discretised regression. Classifiers were trained with either default one-hot encoded labels, or class relationship-aware Gaussian label smoothing with $\sigma = (0.5, 1, 2, 4)$.

Detection

Detection was performed on full-frame samples, using a YOLOv4 network pretrained on the COCO dataset with an input resolution of $800 \times 800 \times 3$ (see Fig 8 F). The standard YOLOv4 training pipeline implemented in darknet was used, with anchors adjusted to allow for detections of both small and large individuals in the same frame [53]. As for classification, the class granularity was either 20 or 5 (see Fig 8 H). Networks were trained for a total of 40,000 iterations, with a decrease in learning rate after 32,000 and 36,000. Although it is best practice to train for at least as many iterations as there are unique training samples (or $\geq$ 2000 iterations per class), loss usually plateaued after approximately 30,000 iterations.

As the out-of-distribution test datasets A and B consisted entirely of cropped frames, the network input resolution was decreased to $128 \times 128 \times 3$. Where the cropped image contained multiple individuals, we used the class prediction with the highest confidence, within 10% of the image centre, as the networks' classification output (Fig. 8 I).

Across all three inference approaches, a total of 98 networks were trained and evaluated.

Performance evaluation

Ideal mass-estimation is accurate, precise and unbiased. Qualitatively accurate networks retain body mass rank order across individuals, and quantitatively accurate networks predict absolute body masses that are close to the ground truth, i. e. they have a small relative error; precise networks provide consistent predictions across different images of the same individual; and unbiased networks have an accuracy and precision that does not vary with ground truth body mass.

Accuracy

Prediction accuracy was assessed on the body mass predictions averaged across all frames of the same individual. The appropriate metric of central tendency differs between regressors, which output continuos variables, and classifiers, which return categorical variables. To account for this difference, regressor accuracy is assessed on arithmetic means, and classification accuracy is assessed on modes.

Qualitative accuracy is assessed through the correlation between the rank order of estimated vs ground truth body masses, appropriately quantified via Spearman's Rank Correlation Coefficient (SRCC). The SRCC is defined as the Pearson correlation coefficient φ between the rank variables $R(X)$ and $R(Y)$:

$$SRCC = \varphi(R(X), R(Y)) = \frac{cov(R(X), R(Y))}{\sigma R(x) \sigma R(Y)} \tag{3}$$

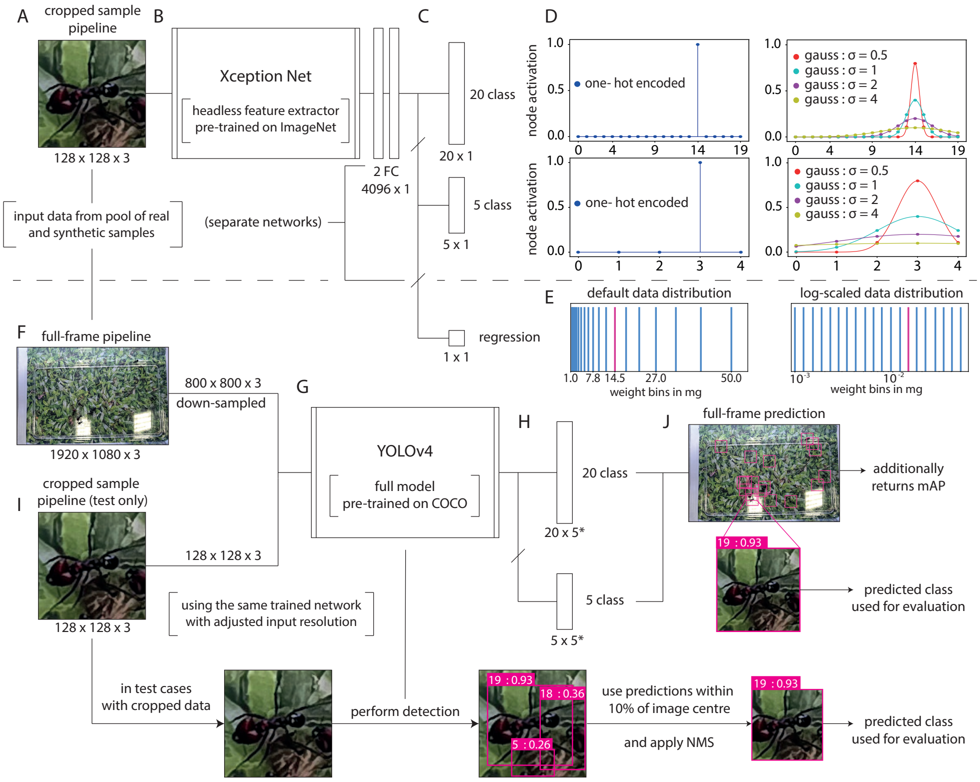


Fig. 8 Schematic overview of deep neural network architectures and training paradigms. The top half of the figure depicts cropped-frame classification and regression; the bottom half shows the information flow in the integrated detector. (A) 128×128 pixel cropped-frame samples were extracted from annotated images, real or synthetic, and fed into (B) a headless Xception Net, pre-trained on the ImageNet dataset. Two fully connected 4,096 node layers followed, and fed into (C) either one of two classifier or a regressor head. The heads and previous fully connected layers were trained individually, using cross entropy-loss for the classifier, and Mean Squared Error (MSE) for the regressor (see Supplementary Table S2 for details). (D) The relationships between classes at training time were encoded by assigning either default one-hot encoded, or custom class-aware Gaussian label smoothing to the classifier's output layer. Label smoothing introduces a differential penalty for misclassification as a function of the distance between ground truth and assigned class, and thus penalises assigning a class 1 worker into class 5 more than assigning the same worker into class 2. (E) Output activations for the regressor were normalised at training time, and labels were log10-transformed to minimise relative error. (F) Full-frame inference was conducted with a detector that returned both the image bounding box and class of individuals simultaneously. The original frames were down-sampled to 800×800 pixel and passed to a YOLOv4 network. The network was pre-trained on the COCO dataset and (H) two different levels of class granularity were trained to retrieve (J) detections at both full-frame and cropped resolution. (I) At test time, the network trained on full-frames can also be used on cropped samples (such as in Test A and Test-B, see 1), by lowering the network input resolution to fit the cropped sample. Only detections within 10% of the image centre were considered, and non-maximum suppression was used to retrieve only the class with the highest activation prediction.

Here, X and Y are the absolute ground truth and predicted body masses, respectively, $cov(R(X), R(Y))$ is the covariance, and $\sigma R(x)$ and $\sigma R(Y)$ are the standard deviations of the rank variables, respectively. The SSRC can fall anywhere between [-1, 1]; a SRCC of unity implies perfect qualitative accuracy, a SRCC of zero implies no association between ranks, and negative unity indicates perfect inverse association.

Quantitative accuracy is assessed with the Mean Absolute Percentage Error (MAPE; also sometimes referred to as Mean Absolute Percentage Deviation (MAPD)):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}(\frac{|y_i - \hat{y}_i|}{\hat{y}_i}) \tag{4}$$

Here, n is the sample size, y_i is the estimated mass, and $\hat{y}_i$ is the ground truth mass. A regressor with perfect quantitative accuracy scores a MAPE of 0, and misclassifying a 1 mg as a 1.5 mg worker results in a MAPE of 50%—the same as misclassifying a 10 mg as a 15 mg worker. However, classifying a 1 mg worker as a 10 mg worker is associated with a MAPE ten times larger than classifying a 10 worker as a 1 mg worker. From these examples thus emerges a caveat that requires comment: because the MAPE quantifies relative instead of absolute errors, it penalises asymmetrically. As a consequence, unless the networks achieve high prediction confidence, they may learn to favour the prediction of small over large body masses, so leading to prediction bias and ultimately potentially even model collapse [57, 59, 60]; this is the main reason that MAPE was not used as a loss function during training, though training on log10-transformed data likely suffers from the same problem. Because the MAPE is defined with respect to the ground-truth value, but classification only returns class centres, classifiers carry an unavoidable inherent error associated with mass discretisation: a classifier with perfect accuracy does not achieve a MAPE of zero, but a MAPE that depends on the distribution of ground truth weights within each class. For the Multi-CamAnts dataset, this error, referred to as $MAPE_{ideal}$, is 6.14% and 22.75% in the validation data for 20-class and 5-class inference approaches, respectively. The natural classifier accuracy metric is not a MAPE, but the categorical classification accuracy: the ratio between the number of correct classifications divided by the total number of classifications. In other words, all mis-classifications are treated identically, regardless of the relative error they carry. Performance evaluation then depends on class number: A 20-class classifier with 20% accuracy classifies every 5th sample correctly and is thus five times better than random class allocation; a 5-class classifier with the same accuracy, in turn, is no better than a random guess. In order to compare the categorical accuracy of classifiers and regressors, the regressor output was translated into a 20-class classification output by assigning each prediction the closest equivalent class centre in linear space. In the interest of simplicity, categorical accuracy is only reported in the supplementary information.

Precision

Network precision is assessed as the variation of weight predictions across frames of the same individual; it is thus also a metric for the stability of predictions. For regressors, prediction precision is assessed through the coefficient of variation (CoV):

$$CoV(Y) = \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma_i}{\mu_i} \tag{5}$$

here, Y_{i_j} are the j predictions for individual i, and σ_i and μ_i are the standard deviation and mean respectively. For direct comparison to a classifier, the mean is replaced by the mode prediction. Note, however, that as the standard deviation is dependent on the mean, this provides a less than ideal ground for direct comparison.

For classifiers, an equivalent metric is not obvious, and it would be most natural to quantify precision via the categorical accuracy defined above, but this time evaluated on a per-individual basis instead of on prediction modes. As a simple means of gauging classifier Prediction Stability (PS) independently of classifier accuracy, we quantify the precision not as the number of correctly predicted classes divided by the number of all samples, but as the average number of samples assigned to the prediction mode $\tilde{y_i}$, divided by all predictions m_i of the same individual i across frames j. Note, however, that as for a direct accuracy comparison, class granularity affects the lower precision bound and inflates results for coarser classifiers.

$$PS(Y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m}p_{ij} \tag{6}$$

$$p_{ij} = \begin{cases} 1 & \text{if } y_{ij} = \tilde{y_i} \\ 0 & \text{if } y_{ij} \neq \tilde{y_i} \end{cases} \tag{7}$$

Bias

There is no *a priori* reason to assume that networks are unbiased, i. e., that their accuracy is independent of the ground truth mass. To quantify network bias, we evaluated the SRCC between the ground truth mass and the MAPE. $SRCC_{acc}$ scores close to zero indicate a predictor with unbiased accuracy; values close to 1 or -1, in turn, indicate a systematic increase or decrease in prediction accuracy with ground truth mass. To provide an accuracy-bias hybrid metric, we further evaluated the coefficient, R^2, with respect to the parity line, $y_p = y_t$:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)}{\sum(y_i - \bar{y}_i)} \tag{8}$$

Unless stated otherwise, the R^2 is reported on log10-scaled data. A perfect network scores an R^2 of unity; any value between [0, 1] indicates that either the network lacks accuracy, carries bias, or both; and a negative value indicates that the network predictions is worse than the sample mean. We do however only report these metrics in the appendices, due to the added difficulty of interpretation and refrain from making comparative claims in the main text.

Human performance

In order to provide an approximate performance baseline, we asked both colleagues with and without experience in leaf-cutter ant research to participate in a weight estimation study. A total of 14 people participated in this exercise; all results were anonymised (See Supplementary Table S3).

A simple online survey was designed to measure human performance on the 5-class cropped-frame weight estimation task (Fig. 9); the survey was implemented in SoSciSurvey [61]). Participants were first briefed on the purpose of the study and the upcoming task; they then agreed to the study conditions, and self-declared whether they work regularly with leaf-cutter ants. Next, participants were shown a simple task description. Akin to the way Wilson [14] used a physical lookup table, participants were shown a digital lookup table as a guide (Fig. 7 and Fig. S 9 B-D). Every participant was initially shown 20 training examples in randomised order; after providing an answer, the correct size class was revealed. The 20 images were sampled from each size class of the original MuliCamAnts training split (see Fig 1 A-C), ensuring that each class was represented equally. After the training phase followed a test phase, during which randomly sampled cropped-frames from the MuliCamAnts validation split, and from Test A and Test B were shown—10 from each dataset for a total of 30 test samples. At this stage, no further feedback was provided, and all answers were recorded for later evaluation. In the evaluation of human performance, we distinguish only between experts and non-experts, i. e. we neglect potential differences in performance on different test data sets, due to sample size restrictions.

A full split of all dataset combinations, network training strategies, and comprehensive performance evaluation on all validation and test data is provided in **Supplementary table 2**.

References

[1] Hamdan, M.K.A., Just, J.: Mass Estimation from Images using Deep Neural Network and Sparse Ground Truth (2019)

[2] Ponce, J.M., Aquino, A., Millan, B., Andujar, J.M.: Automatic Counting and Individual Size and Mass Estimation of Olive-Fruits Through Computer Vision Techniques. IEEE Access **7**, 59451–59465 (2019) https://doi.org/10.1109/ACCESS.2019.2915169

[3] Standley, T., Sener, O., Chen, D., Savarese, S.: image2mass: Estimating the mass of an object from its image. In: Levine, S., Vanhoucke, V., Goldberg, K. (eds.) Proceedings of the 1st Annual Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 78, pp. 324–333. PMLR, ??? (2017). https://proceedings.mlr.press/v78/standley17a.html

[4] G. Vivek Venkatesh, A.G. S. Md. Iqbal, Ganesan, D.: Estimation of volume and mass of axi-symmetric fruits using image processing technique. International Journal of Food Properties **18**(3), 608–626 (2015) https://doi.org/10.1080/10942912.2013.831444 https://doi.org/10.1080/10942912.2013.831444

[5] Suwannakhun, S., Daungmala, P.: Estimating Pig Weight with Digital Image Processing using Deep Learning. Proceedings - 14th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2018 (5), 320–326 (2018) https://doi.org/10.1109/SITIS.2018.00056

[6] Gjergji, M., De Moraes Weber, V., Otávio Campos Silva, L., Da Costa Gomes, R., De Araújo, T.L.A.C., Pistori, H., Alvarez, M.: Deep Learning Techniques for Beef Cattle Body Weight Prediction. Proceedings of the International Joint Conference on Neural Networks (2020) https://doi.org/10.1109/IJCNN48605.2020.9207624

[7] Andrade, J.M.L., Moreno, P.: Improving the Estimation of Object mass from images. 2023 IEEE International Conference on Autonomous Robot Systems and Competitions, ICARSC 2023, 199–206 (2023) https://doi.org/10.1109/ICARSC58346.2023.10129573

[8] Nir, O., Parmet, Y., Werner, D., Adin, G., Halachmi, I.: 3D Computer-vision system for automatically estimating heifer height and body mass. Biosystems Engineering **173**, 4–10 (2018) https://doi.org/10.1016/j.biosystemseng.2017.11.014

[9] Dohmen, R., Catal, C., Liu, Q.: Image-based body mass prediction of heifers using deep neural networks. Biosystems Engineering **204**, 283–293 (2021) https://doi.org/10.1016/j.biosystemseng.2021.02.001

[10] Hu, C., Kong, S., Wang, R., Zhang, F., Wang, L.: Insect mass estimation based on radar cross section parameters and support vector regression algorithm. Remote

Sensing **12**(11), 1–11 (2020) https://doi.org/10.3390/rs12111903

[11] Hu, C., Zhang, F., Li, W., Wang, R., Yu, T.: Estimating Insect Body Size From Radar Observations Using Feature Selection and Machine Learning. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2022) https://doi.org/10.1109/TGRS.2022.3224618

[12] Eder, E.B., Almonacid, J.S., Delrieux, C., Lewis, M.N.: Body volume and mass estimation of southern elephant seals using 3D range scanning and neural network models. Marine Mammal Science **38**(3), 1037–1049 (2022) https://doi.org/10.1111/mms.12910

[13] Wetterer, J.K.: Allometry and the geometry of leaf-cutting in Atta cephalotes. Behav Ecol Sociobiol **29**(5), 347–351 (1991)

[14] Wilson, E.O.: Caste and division of labor in leaf-cutter ants (Hymenoptera: Formicidae :Atta). I. The overall pattern in A. sexdens. Behav Ecol Sociobiol **7**(2), 143–156 (1980)

[15] Wilson, E.O.: Caste and division of labor in leaf-cutter ants Hymenoptera: Formicidae: Atta). III. Ergonomic resiliency in foraging by Atta cephalotes. Behav Ecol Sociobiol **14**, 47–54 (1983)

[16] Wetterer, J.K.: The Ecology and Evolution of Worker Size-Distribution in Leaf-Cutting Ants (Hymenoptera: Formicidae). Sociobiology **34**, 119–144 (1999)

[17] Wilson, E.O.: Caste and Division of Labor in Leaf-Cutter Ants (Hymenoptera : Formicidae : Atta) **156** (1980)

[18] Ferguson-Gow, H., Sumner, S., Bourke, A.F.G., Jones, K.E.: Colony size predicts division of labour in attine ants. Proc R Soc B **281**(1793), 20141411 (2014)

[19] Hölldobler, B., Wilson, E.O.: The ants. Harvard University Press (1990)

[20] Hölldobler, B., Wilson, E.O.: The leafcutter ants: civilization by instinct. WW Norton & Company (2010)

[21] Wilson, E.O.: Caste and division of labor in leaf-cutter ants (Hymenoptera: Formicidae: Atta.): II. The Ergonomic Optimization of Leaf Cutting. Behav Ecol Sociobiol **7**(2), 143–156 (1980)

[22] Wilson, E.O.: Caste and division of labor in leaf-cutter ants (Hymenoptera: Formicidae: Atta.) IV. Colony ontogeny of A. cephalotes. Behav Ecol Sociobiol **14**(1), 47–54 (1983)

[23] Wilson, E.O.: Caste and division of labor in leaf-cutter ants (Hymenoptera: Formicidae:, 47–54 (1983)

[24] Wilson, E.O.: The origin and evolution of polymorphism in ants. The Quarterly Review of Biology **28**(2), 136–156 (1953)

[25] Wilson, E.O.: Caste and division of labor in leaf-cutter ants The colonies were collected at the earliest stages of development, 55–60 (1983)

[26] Roces, F., Hölldobler, B.: Use of stridulation in foraging leaf-cutting ants: mechanical support during cutting or short-range recruitment signal? Behav Ecol Sociobiol **39**(5), 293–299 (1996)

[27] Hoelldobler, B.: Territorial behavior in the green tree ant (Oecophylla smaragdina). Biotropica, 241–250 (1983)

[28] Wetterer, J.K.: Allometry and the geometry of leaf-cutting in Atta cephMotes, 347–351 (1991)

[29] Wetterer, J.K.: Ontogenetic changes in forager polymorphism and foraging ecology in the leaf-cutting ant Atta cephalotes. Oecologia **98**(2), 235–238 (1994)

[30] Wilson, E.O.: The ergonomics of caste in the social insects. Am Nat **102**(923), 41–66 (1968)

[31] Oster, G.F., Wilson, E.O.: Caste and ecology in the social insects. Princeton University Press (1978)

[32] Clark, E.: Dynamic matching of forager size to resources in the continuously polymorphic leaf-cutter ant, Atta colombica (Hymenoptera, Formicidae). Ecol Entomol **31**(6), 629–635 (2006)

[33] Helanterä, H., Ratnieks, F.L.W.: Geometry explains the benefits of division of labour in a leafcutter ant. Proceedings of the Royal Society of London B: Biological Sciences **275**(1640), 1255–1260 (2008)

[34] Püffel, F., Roces, F., Labonte, D.: Strong positive allometry of bite force in leaf-cutter ants increases the range of cuttable plant tissues. Journal of Experimental Biology **226**(13) (2023) https://doi.org/10.1242/jeb.245140

[35] Burd, M.: Variable load size-ant size matching in leaf-cutting ants,Atta colombica (Hymenoptera: Formicidae). Journal of Insect Behavior **8**(5), 715–722 (1995) https://doi.org/10.1007/BF01997240

[36] Billick, I.: The relationship between the distribution of worker sizes and new worker production in the ant Formica neorufibarbis. Oecologia **132**(2), 244–249 (2002) https://doi.org/10.1007/s00442-002-0976-7

[37] Imirzian, N., Puffel, F., Labonte, D.: 3d shape analysis of polymorphic leafcutter ant mandibles. In: INTEGRATIVE AND COMPARATIVE BIOLOGY, vol. 62, pp. 152–152 (2023). OXFORD UNIV PRESS INC JOURNALS DEPT, 2001

EVANS RD, CARY, NC 27513 USA

[38] Püffel, F., Pouget, A., Liu, X., Zuber, M., Van De Kamp, T., Roces, F., Labonte, D.: Morphological determinants of bite force capacity in insects: A biomechanical analysis of polymorphic leaf-cutter ants. Journal of the Royal Society Interface **18**(182) (2021) https://doi.org/10.1098/rsif.2021.0424

[39] Hernández, J.: Charaterization of the mandible and mandibular glands in different castes of the leaf-cutting ant atta laevigata (f. smith)(hymenoptera: Formicidae) using scanning electron microscopy. Bol. Entomol. Venez. NS **10**, 51–56 (1995)

[40] Silva, L.C., Camargo, R.S., Lopes, J.F.S., Forti, L.C.: Mandibles of Leaf-Cutting Ants: Morphology Related to Food Preference. Sociobiology **63**(3), 881–888 (2016)

[41] Feener, D.H., Lighton, J.R.B., Bartholomew, G.A.: Curvilinear Allometry, Energetics and Foraging Ecology: A Comparison of Leaf-Cutting Ants and Army Ants. Functional Ecology **2**(4), 509 (1988) https://doi.org/10.2307/2389394

[42] Feener, D.H., Lighton, J.R.B., Bartholomew, G.A.: Curvilinear allometry, energetics and foraging ecology: a comparison of leaf-cutting ants and army ants. Functional Ecology, 509–520 (1988)

[43] Muratore, I.B., Ilieş, I., Huzar, A.K., Zaidi, F.H., Traniello, J.F.A.: Morphological evolution and the behavioral organization of agricultural division of labor in the leafcutter ant Atta cephalotes. Behavioral Ecology and Sociobiology **77**(6), 70 (2023) https://doi.org/10.1007/s00265-023-03344-4

[44] Plum, F., Labonte, D.: scAnt —an open-source platform for the creation of 3D models of arthropods (and other small objects) . PeerJ **9**, 11155 (2021) https://doi.org/10.7717/peerj.11155

[45] Plum, F., Bulla, R., Beck, H., Imirzian, N., Labonte, D.: replicAnt - generating annotated images of animals in complex environments with Unreal Engine. Nat Commun. (accepted), 2023–0420537685 (2023) https://doi.org/10.1101/2023.04.20.537685

[46] Plum, F.: Omnitrax: A deep learning-driven multi-animal tracking and pose-estimation add-on for blender. Journal of Open Source Software **9**(95), 5549 (2024) https://doi.org/10.21105/joss.05549

[47] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 **2017-January**, 1800–1807 (2017) https://doi.org/10.1109/CVPR.2017.195 arXiv:1610.02357

[48] Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–15 (2015) arXiv:1412.6980

[49] Hendrycks, D., Lee, K., Mazeika, M.: Using pre-training can improve model robustness and uncertainty. 36th International Conference on Machine Learning, ICML 2019 **2019-June**(2018), 4815–4826 (2019) arXiv:1901.09960

[50] Krizhevsky, A., Sutskever, I., Hinton., G.E.: Imagenet classification with deep convolutional neural networks. In Advances in neural information. Advances in neural information processing systems, 1097–1105 (2012) arXiv:1102.0183

[51] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8693 LNCS**(PART 5), 740–755 (2014) https://doi.org/10.1007/978-3-319-10602-1_48 arXiv:1405.0312

[52] Krizhevsky, A.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)

[53] Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: Optimal Speed and Accuracy of Object Detection (2020) arXiv:2004.10934

[54] Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., Papadoyannis, E.S., Normand, E., Deutsch, D.S., Wang, Z.Y., McKenzie-Smith, G.C., Mitelut, C.C., Castro, M.D., D'Uva, J., Kislin, M., Sanes, D.H., Kocher, S.D., Wang, S.S.H., Falkner, A.L., Shaevitz, J.W., Murthy, M.: SLEAP: A deep learning system for multi-animal pose tracking. Nature Methods **19**(4), 486–495 (2022) https://doi.org/10.1038/s41592-022-01426-1

[55] Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D.: Fast and robust animal pose estimation. bioRxiv, 620245 (2019) https://doi.org/10.1101/620245

[56] Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M.M., Di Santo, V., Soberanes, D., Feng, G., Murthy, V.N., Lauder, G., Dulac, C., Mathis, M.W., Mathis, A.: Multi-animal pose estimation, identification and tracking with DeepLabCut. Nature Methods **19**(4), 496–504 (2022) https://doi.org/10.1038/s41592-022-01443-0

[57] Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ. Computer science **7**, 623 (2021) https://doi.org/10.7717/peerj-cs.623

[58] Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)

[59] Tofallis, C.: A better measure of relative prediction accuracy for model selection and model estimation. Journal of the Operational Research Society **66**(8), 1352–1362 (2015) https://doi.org/10.1057/jors.2014.103 arXiv:2105.05249

[60] Makridakis, S.: Accuracy measures: theoretical and practical concerns. International Journal of Forecasting **9**(4), 527–529 (1993)

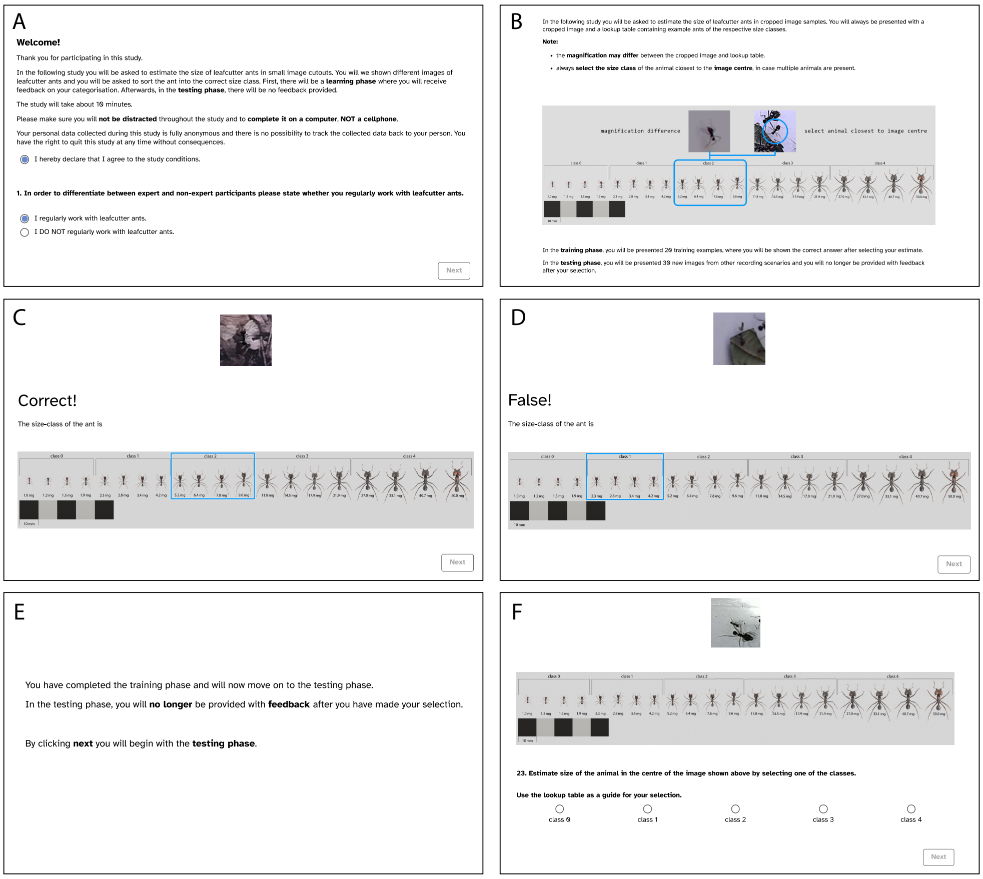
[61] Leiner, D.J.: SoSci Survey (Version 3.1.06) (2019). https://www.soscisurvey.de

Acknowledgements

This study was funded by the Imperial College's President's PhD Scholarship (to Fabian Plum) and is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 851705, to David Labonte). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would like to thank Theo Humbeeck, who aided in the collection and annotation of the out-of-distribution dataset *Test A*.

Supplementary information

Survey - Human weight estimation



Welcome!

Thank you for participating in this study.

In the following study you will be asked to estimate the size of leafcutter ants in small image cutouts. You will we shown different images of leafcutter ants and you will be asked to sort the ant into the correct size class. First, there will be a **learning phase** where you will receive feedback on your categorisation. Afterwards, in the **testing phase**, there will be no feedback provided.

The study will take about 10 minutes.

Please make sure you will **not be distracted** throughout the study and to **complete it on a computer**, **NOT a cellphone**.

Your personal data collected during this study is fully anonymous and there is no possibility to track the collected data back to your person. You have the right to quit this study at any time without consequences.

● I hereby declare that I agree to the study conditions.

1. In order to differentiate between expert and non-expert participants please state whether you regularly work with leafcutter ants.

● I regularly work with leafcutter ants.
○ I DO NOT regularly work with leafcutter ants.

In the following study you will be asked to estimate the size of leafcutter ants in cropped image samples. You will always be presented with a cropped image and a lookup table containing example ants of the respective size classes.

Note:

- the **magnification may differ** between the cropped image and lookup table.
- always **select the size class** of the animal closest to the **image centre**, in case multiple animals are present.

In the **training phase**, you will be presented 20 training examples, where you will be shown the correct answer after selecting your estimate.

In the **testing phase**, you will be presented 30 new images from other recording scenarios and you will no longer be provided with feedback after your selection.

Correct!

The size-class of the ant is

False!

The size-class of the ant is

You have completed the training phase and will now move on to the testing phase.

In the testing phase, you will **no longer** be provided with **feedback** after you have made your selection.

By clicking **next** you will begin with the **testing phase**.

23. Estimate size of the animal in the centre of the image shown above by selecting one of the classes.

Use the lookup table as a guide for your selection.

○ class 0 ○ class 1 ○ class 2 ○ class 3 ○ class 4

Fig. 9 Main pages of a survey designed to measure human performance on 5-class weight estimation tasks. (A) Welcome page introducing the participant to the study and prompting them to agree to the study conditions, as well as declare themselves either an expert or non-expert by selecting whether or not they regularly work with leaf-cutter ants. (B) Task description: This page outlines the weight estimation task, explaining what type of images will be prompted, and how the participant is supposed to enter their answer; it also explains the differences between the training and testing phases of the survey. (C-D) Example pages from the training phase after the participant has made a correct or incorrect selection respectively. (E) This page is prompted after the participant has completed their training. (F) In the testing phase, no further feedback is provided after each estimate has been made.

Augmenting real annotated data with synthetic samples improves network robustness

The large volume of hand-annotated data—up to 2.5 million samples in cropped inference approaches—sufficed to train models that achieved good performance on the validation data. These models consequently benefited only little from the addition of synthetic training data (see 10); in some cases, the inference performance even slightly decreased, likely reflecting "beneficial" overfitting to the within-distribution validation case.

Synthetic data did however make inference more robust (see 10, B,C,H,I): the networks that did best on Out-Of-Distribution (OOD) data were trained on a mixture of real and synthetic data (see Fig. 7; MAPE scores were lower, potentially due to a more favourable distribution of weight classes in the out-of-distribution data. see 10 F as well as section 2). This effect was particularly evident for recordings from cluttered environments (Fig. 10, B,C,H,I), and mirrors similar results in earlier work, which suggested that synthetic data can help to embed a subject-specific understanding into the networks [45]. In the context of body mass inference, a key strength of synthetic data is that it can be generated such that size-related differences in cropped image occupancy or compression artefacts are entirely avoided, so preventing networks from learning to infer size from these artefactual features, which would impede generalisation.

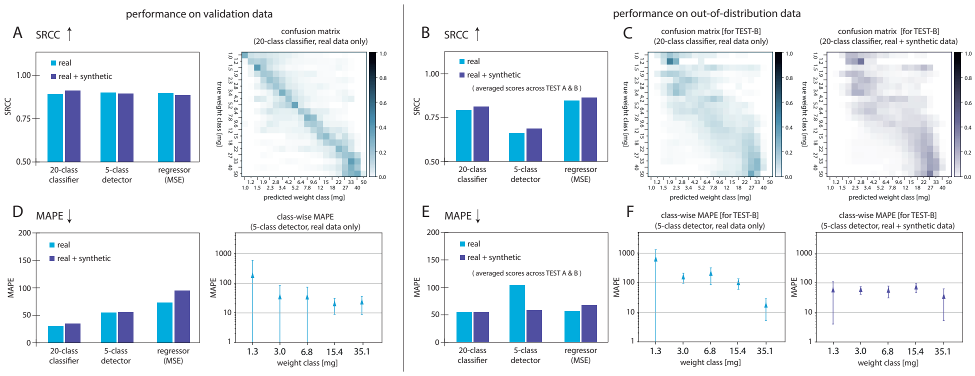


Fig. 10 **Augmenting training data with synthetic samples improved network robustness.** Representative examples of different weight inference strategies, including a 20-class classifier (no label smoothing), a regressor trained with MSE loss, and a 5 class detector, which all benefited from synthetic data. The addition of synthetic data had only small effects on performance on validation data (A,D,G), but their strength became apparent in out-of-distribution examples. (A) Accuracy of networks on unseen validation samples (see 1 A), and a confusion matrix for the 5-class detector. (B) Synthetic data increased accuracy for all networks on out-of-distribution data, and most notably for the detector. (C) Confusion matrices of the detector on Test-B footage, trained without (left) and with (right) synthetic data. (D) MAPE scores of networks on unseen validation samples and the class-wise MAPE scores of the 20-class classifier on the same data (lower is better). (E) MAPE scores on out-of-distribution test data changed little upon the addition of synthetic data. (F) Class-wise MAPE scores on Test-B with the 20-class classifier remain elevated. (G) R^2 scores of networks on unseen validation samples (see 1 A), and a parity plot showing the grouped predicted vs ground truth weight of the regressor for all individuals. (H) The R^2 score is elevated in all examples, indicating that the addition of synthetic data brings the estimates closer to the respective ground truth values in unseen conditions. (I) Prediction vs ground truth plots of the regressor for Test A data.

Gaussian label smoothing can improve network precision and robustness.

One hot-encoded labels are a sensible choice for categorical inference tasks. However, classes in discretised weight inference retain an ordinal characteristic, so that not all classification errors are equal. Consequently, penalising incorrect classification into adjacent classes less can avoid overfitting, and increases the correlation between performance metrics in out-of-distribution data (see table 1). The 5-class classifier, trained on mixed MultiCamAnts and synth-simple data with class relationship-aware Gaussian label smoothing with $\sigma = 2$, achieved the highest overall accuracy of 47.3% on out-of-distribution data. There also appeared to be a systematic decrease in the CoV, indicating increased classification precision. The best smoothing parameter σ varied with the number of classes; for 5 class models, the activation profile became flat, and performance in fact decreased for $\sigma > 2$ (see Table 1).

Coarser classification approaches outperform deep regressors.

Classifiers and detectors usually outperformed regressors, even in regression centred evaluation metrics such as MAPE or R^2 (see table 2). Careful comparison of classifier

Table 1 Performance of 5 and 20 class classifiers trained with class relationship-aware Gaussian label smoothing. All networks were trained with mixed datasets containing the default real 1 (A-C) and the simple synthetic dataset 7 (A-C), comprising 2.5 million and 910,000 samples respectively. MAPE, accuracy, Coefficient of Variation (CoV), and R^2 are reported on validation data (500,000 unseen samples collected across the camera perspectives and background textures depicted in 1 (A-C)), and on out-of-distribution datasets A and B, comprising 4,944 and 30,526 samples respectively (see 1 (D-I).) Label smoothing increased both robustness and precision, as evident in increased performance metrics on out of distribution data, and a reduction of the CoV.

		Validation scores				combined scores on Test A and B			
classes	sigma	SRCC ↑	MAPE ↓	acc. ↑	CoV ↓	SRCC ↑	MAPE ↓	acc. ↑	CoV ↓
5	0*	0.840	73.3	0.610	0.718	0.732	67.7	0.401	0.560
	0.5	0.872	54.3	0.622	0.714	0.774	61.3	0.410	0.565
	1	0.923	42.2	0.595	0.632	0.738	55.2	0.424	0.553
	2	0.903	53.0	0.544	0.515	0.626	60.5	0.436	0.480
	4	0.851	78.6	0.453	0.436	0.444	70.8	0.415	0.413
20	0*	0.888	41.9	0.420	0.784	0.686	56.0	0.105	0.622
	0.5	0.885	39.5	0.413	0.776	0.590	54.0	0.133	0.545
	1	0.861	56.6	0.383	0.717	0.714	60.3	0.115	0.608
	2	0.872	37.8	0.319	0.747	0.653	73.2	0.109	0.533
	4	0.915	36.4	0.241	0.619	0.831	57.2	0.126	0.477

* using default one-hot encoding without label smoothing.

and regressor performance requires to take into account that the baseline performance varies with class number: A 5-class classifier with an accuracy of 40 % performs twice as well as a random guess, but a 20-class classifier with the same performance is eight times better than random.

Table 2 Inference performance with different class granularities. All networks were trained with mixed datasets containing the default real training 1 (A-C) and the simple synthetic dataset 7 (A-C), comprising 2.5 million and 910,000 samples respectively. MAPE, accuracy, Coefficient of Variation, and R^2 are reported on validation data (500,000 unseen samples collected across the camera perspectives and background textures depicted in 1 (A-C)), and averaged performance on out-of-distribution datasets A and B, comprising 4,944 and 30,526 samples respectively (see Fig 1 (D-I).) We find that coarser class granularity often positively affects the resulting performance

	Validation scores				combined scores on Test A and B			
type	SRCC ↑	MAPE ↓	acc. ↑	CoV ↓	SRCC ↑	MAPE ↓	acc. ↑	CoV ↓
CLASS 5 ($\sigma = 1$)	0.923	38.9	0.598	0.611	0.759	55.1	0.446	0.519
CLASS 20 ($\sigma = 4$)	0.929	35.1	0.255	0.615	0.818	56.4	0.131	0.490
REG MSE	0.810	165.4	0.113	0.313	0.743	105.8	0.135	0.297
REG MSE LOG	0.886	95.8	0.135	0.353	0.868	68.0	0.107	0.324
DETECT 5	0.896	57.2	0.708	0.589	0.687	59.4	0.289	0.607
DETECT 20	0.892	60.1	0.546	0.583	0.609	112.2	0.067	0.664

Lower size-classes disproportionately affect MAPE scores

Regardless of inference approach, loss function, and label transformation technique employed, lower size classes disproportionately affect the overall MAPE scores, as evident from inspection of the class-wise MAPE (see 11, as well as appendix 2), which was typically between 3 to 10 times higher for the smallest classes, even for overall well performing inference approaches.

In addition to the size-dependence inherent in the definition of the MAPE score (see methods), the presence of larger individuals occluding the target animal in the same cropped frame likely inflates the error further. If inference approaches are selected according to the lowest MAPE, then a method that systematically underestimates body mass will do better than a method that systematically overestimates it: the MAPE favours models that underpredict the target distribution because it assigns more weight to data points with smaller ground truth values in the denominator, making these points more influential [59, 60]. Various additions have been suggested to counter act this property such as dividing the absolute error by the average of the predicted and ground truth value instead of the ground-truth alone [60] or by log-transformation of the MAPE [59], and may be explored in future work.

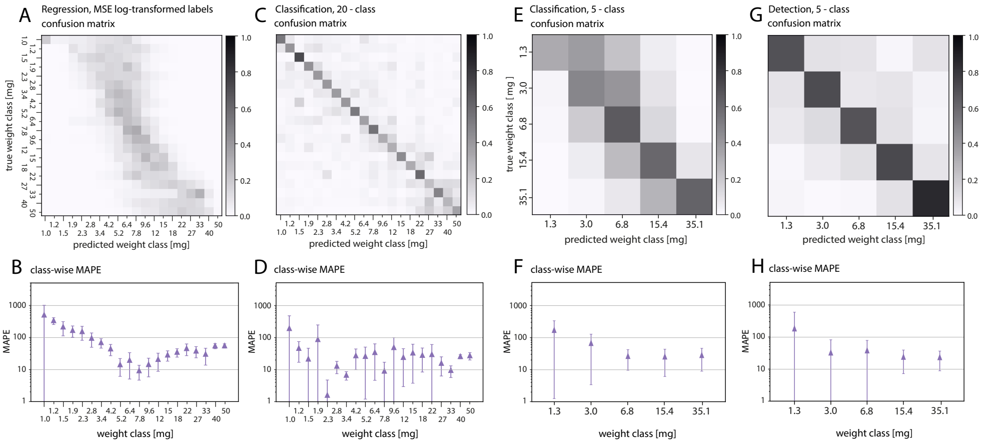


Fig. 11 **Weight estimation performance of classifiers, regressors and detectors.** Accuracy on unseen validation data increases from left to right; all networks were trained on a mix of real and synthetic data. (A) A regressor, trained with log-transformed labels achieved a categorical accuracy of 0.135; (C) The 20-class classifier trained with class relationship-aware Gaussian label smoothing ($\sigma = 0.5$), achieved an accuracy of 0.420; (E) A 5-class classifier, trained with class relationship-aware Gaussian label smoothing ($\sigma = 2$), achieved an accuracy of 0.538; and (G) a 5-class detector, trained with default labels, achieved an accuracy of 0.708. The associated MAPE scores are nevertheless high, at (B) 95.8; (D) 25.9; (F) 59.9; and (H) 57.2 respectively, likely because identical absolute weight errors lead to large relative errors for small weight classes, which skew the MAPE score