JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Leakage-Aware Cross-Dataset Multi-Task Learning for Shrimp Disease Recognition and Biomass Estimation

Budi Juarto * and Mochamad Haldi Widianto

Computer Science Department, Bina Nusantara University, Jakarta 11530, Indonesia;
e-mail: budi.juarto@binus.ac.id; mochamad.widianto@binus.ac.id
* Corresponding Author : Budi Juarto

Abstract: Public shrimp image collections are reused to train disease classifiers and non-invasive biomass estimators, yet they mix originals with published augmentations, byte-identical copies, and repeated views of the same animal, potentially inflating performance through data leakage. This study examines what these collections support once leakage is controlled: how protocol and acquisition conditions, rather than architecture alone, influence performance; whether disease recognition transfers to an independently collected dataset; and whether monocular imagery supports individual-weight estimation. One encoder is trained across three public datasets under masked supervision, using a grouping strategy that accounts for duplicates, augmentations, near-duplicates, and repeated views; a 22-check audit guides training and evaluation, with 240 runs reported. Regrouping by animal rather than file eliminates the 84.8–92.6% animal overlap in ShrimpDiseaseImageBD test folds, reducing it to 0.0%. A random image-level split yields 0.976 balanced accuracy, but 63% of its test records are published augmented copies; using originals alone yields 0.947, compared with 0.934 in the augmentation-enabled group-aware arm, which tests different original records. Transfer is asymmetric: the best-performing encoder reaches 0.615 balanced accuracy and 0.751 area under the curve on TigerShrimpBD; on ShrimpDiseaseImageBD, every encoder ranks below chance, with values of 0.271–0.390. TigerShrimpBD's 0.998 within-dataset score is affected by acquisition confounding, as every recovered capture session contains only one health class. Holding out a physical site increases biomass error from 3.18 g to 9.20 g, indicating limited cross-site generalization. The contribution is a reusable audit and evaluation protocol for interpreting performance on such collections.

Keywords: Acquisition Confounding; Biomass Estimation; Data Leakage; Disease Recognition; Domain Generalization; Multi-Task Learning; Partial Labels; Shrimp Aquaculture.

Received: August, 10th 2026
Revised: October, 5th 2026
Accepted: October, 6th 2026
Published: November, 9th 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licenses (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Shrimp aquaculture is economically important but remains highly vulnerable to disease outbreaks. White Spot Syndrome Virus (WSSV), for example, has been associated with economic losses amounting to billions of dollars and can cause an entire pond to be lost within days [1]. In this context, farmed-shrimp photographs have increasingly been considered a potential source of information for non-invasive monitoring. Visible disease manifestations, such as white calcareous spots on the carapace, melanised gill tissue in the branchial chamber, and a yellowish cephalothorax, can be captured using conventional cameras under suitable imaging conditions. Biomass estimation, however, presents a different challenge. Although body shape and proportions can be observed in a photograph, absolute body size cannot necessarily be recovered without an appropriate scale reference. Linear body measurements are commonly used to characterise shrimp size, while body weight is generally related to these measurements through an allometric relationship [2]. Consequently, disease recognition and biomass estimation from images involve different sources of uncertainty and require evaluation strategies that account for their respective limitations.

Despite the growing use of image-based methods for shrimp analysis, most published studies have relied on evaluations within individual datasets, leaving their generalisation under

different data conditions insufficiently established. Several recurring evaluation challenges can affect the reliability of reported performance. First, image collections may contain records that are not statistically independent. Published disease-image studies use different evaluation designs: some cross-validate original and augmented datasets [3], whereas others use random image-level folds without reporting individual-level grouping [4]. If related records cross partitions, information overlap can inflate performance estimates [5], [6]. Second, acquisition-related characteristics may become correlated with the target labels. When differences in capture sessions, imaging conditions, or image characteristics are associated with particular classes, models may exploit these cues rather than learn disease-related visual features [7]. In fake-news detection, a verifiability-focused review similarly cautions that near-perfect reported scores may reflect leakage or corpus separability rather than genuine generalization [8]. Therefore, performance measured under controlled or collection-specific acquisition conditions cannot automatically be assumed to generalise to images captured under different conditions at farm level. Third, the available datasets do not share identical annotation schemes. ShrimpDiseaseImageBD [9], TigerShrimpBD [10] and a Litopenaeus vannamei biomass collection [11] provide different types of supervision. Consequently, unannotated attributes cannot be treated as negative labels when these datasets are combined, as doing so would introduce incorrect supervision into the learning process.

These challenges indicate that reliable evaluation of image-based shrimp analysis requires more than selecting an appropriate model architecture. Related records must be grouped to prevent information overlap, differences in label availability must be accommodated during training, and generalisation must be assessed across independently collected datasets. Acquisition-related confounding also requires separate investigation, as it may persist despite leakage-aware data partitioning. For biomass estimation, evaluation should distinguish within-collection performance from generalisation to unseen databases or sites, where acquisition conditions and population characteristics may affect the relationship between image features and individual body weight.

Accordingly, this study investigates whether evaluation protocols and acquisition conditions have a greater influence on the measured performance of image-based shrimp analysis than model architecture alone. Three aspects are examined: the effects of leakage-aware grouping and acquisition confounding on disease recognition, the transferability of disease recognition across independently collected datasets, and the limitations of individual-level biomass estimation from monocular images. A shared-encoder multi-task learning framework with masked supervision is employed to accommodate differences in label availability, while leakage-aware grouping and an acquisition audit support the evaluation process. Within-dataset, cross-dataset, and biomass experiments are conducted to examine performance under different data partitions and collection conditions. The methodological components are used to support this investigation rather than presented as independent architectural contributions.

The contributions of this study are therefore centred on the evaluation and empirical analysis of publicly available shrimp image collections. First, conventional image-level evaluation is contrasted with leakage-aware, group-based evaluation to examine how related records affect measured disease-recognition performance. Second, cross-dataset experiments are conducted in both transfer directions to assess whether disease-recognition performance is maintained across independently collected datasets and to identify potential effects of acquisition differences. Third, individual-level biomass estimation is evaluated within databases and under database- and site-held-out settings to examine the extent to which image-based predictions capture individual weight rather than collection-specific characteristics. The remainder of this paper is organised as follows. Section 2 reviews related work, Section 3 describes the materials and methods, Section 4 presents the results and discussion, Section 5 compares the findings with previously reported results, and Section 6 concludes the study.

2. Related Work

2.1. Vision-Based Shrimp Disease and Biomass Estimation

Image-based shrimp disease recognition reports consistently strong numbers: multi-model ensembles [12], a histopathology pathogen classifier [13], a lightweight model reporting 93% five-fold accuracy over the Black Gill, WSSV, Yellow Head and Healthy classes of the public collections used here [4], and a mobile detector reporting 100% accuracy with originals

and augmentations cross-validated together on a private set [3]. Only [4] reports a test across those collections, in one direction; only the abstract of [12] was available for independent verification. The pattern is not shrimp-specific: in this journal an enhanced vision transformer reports 97.92% rice-disease accuracy [14]; transfer beyond that collection requires separate evaluation. Biomass work takes its spatial scale from the apparatus: feed-tray geometry [15], calibrated stereo [16], morphometric features [17]. Monocular estimation is ill-posed without such a constraint [18], and weight is conventionally modelled from length by a near-cubic allometric law [2].

2.2. Partial Labels, Domain Shift and Evaluation Validity

Partial-label multi-label learning treats unobserved annotations as unknown rather than negative [19]; this principle supports the masked supervision used here. Homoscedastic uncertainty can balance multi-task losses [20]. MixStyle mixes source feature statistics for domain generalization [21], whereas Deep CORAL aligns feature correlations for domain adaptation [22]; our CORAL variant aligns training/source-domain batches. Plant-disease models trained under controlled conditions can lose accuracy in the field [23], and an aquaculture review cautions that favorable research conditions may overestimate commercial performance [24]. Leakage taxonomies describe duplicated records, non-independent splits and contaminated partitions [25]. In molecular data, reducing cross-split similarity can lower measured performance [5]. Medical-imaging models can select acquisition-related shortcuts and fail externally [26]; a systematic review found widespread methodological limitations, including confounded evaluation, that prevented clinical use of the reviewed COVID-19 models [27]. A review of plant-disease datasets likewise warns that confounding and inappropriate random splits can inflate performance [6].

2.3. Research Gap

Table 1 summarises these families, the setting each is validated under, the gap each leaves open, and how this study addresses it.

Table 1. Research gaps identified from previous work, and how this study addresses each of them

Approach	Validated under	Gap	This study
Shrimp disease CNNs [3], [4] and [12]	Within-corpus validation; random or k-fold splits where reported	Original/augmented cross-validation in [3]; individual grouping not reported in [4]; one cross-corpus test [4]	Ladder: random, group-aware, cross-dataset
Histopathology pathogen classifiers [13]	Microscopy patches with slide identifiers	Gross morphology unobserved; farm photographs lack identifiers	Grouping keys recovered from images and metadata
Rig-fixed, stereo or morphometric biomass [15]–[17]	Single site; feed-tray geometry or calibrated stereo	Scale from the apparatus, not image content	Image-only weight across databases and sites, against a measurement oracle
Partially supervised multi-label learning [19]	Curated multi-label benchmarks, partial labels	Duplicate and repeated-subject control is not the focus	Masked losses and leakage-safe grouping enforced jointly
Leakage and shortcut studies elsewhere [6], [25] and [26]	Clinical, agricultural and general scientific data	Never measured on shrimp imagery	Audit gate counts every relation and confound

Taken together, these gaps indicate the need for an integrated evaluation framework that jointly addresses data leakage, acquisition-related confounding, cross-dataset generalization, and the limitations of image-based biomass estimation. Rather than introducing a new network architecture, this study contributes a masked multi-task formulation, a leakage-aware evaluation protocol, and an empirical assessment of what existing public shrimp image collections can reliably support.

3. Materials and Methods

The proposed methodology comprises four components: a partially supervised formulation with masked objectives, a data audit that gates model training, a leakage-safe grouping

and splitting procedure, and a shared-encoder multi-task network. Figure 1 illustrates how these components are integrated. The audit gate evaluates the three public data sources and establishes fused groups before model training. These groups, rather than individual images, serve as the units for all leakage-aware data splits. The task-specific heads are trained using masked objectives, ensuring that labels unavailable in a particular dataset contribute neither to the loss nor to the corresponding gradients. The image counts shown in Figure 1 represent the audited originals in the primary disease pool and the complete biomass collection.

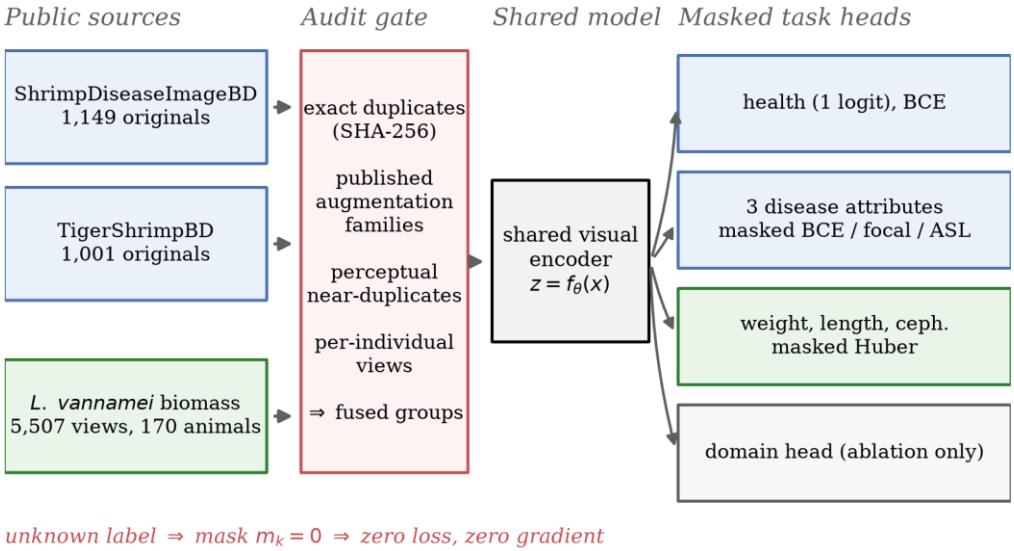


Figure 1. Study pipeline: three public datasets pass an audit gate that assigns fused groups to a shared encoder with four masked heads

3.1. Datasets

Three publicly available datasets released under the CC BY 4.0 license were used in this study. Each dataset was verified against the publisher-declared checksums associated with its DOI. For the biomass collection, Mendeley Data version 2 was used because version 1 could not be retrieved. The datasets, their available supervision, and the corresponding grouping units are summarized in Table 2. Related records and images associated with the same identified animal were assigned to the same group. Targets not annotated by a particular source were masked during training: Yellow Head labels were unavailable throughout ShrimpDiseaseImageBD, while cephalothorax length was unrecorded in DB1. All disease labels considered in this study were available in TigerShrimpBD.

Table 2. Datasets, supervision and splitting units

Dataset	Images / originals	Supervision	Grouping unit	Groups	Role
ShrimpDiseaseImageBD (V3, 10.17632/jhrtdj9txm.3)	1,895 / 1,149	health; Black Gill; WSSV	animal-level group	412	disease
TigerShrimpBD (V1, 10.17632/9dj4sk5d55.1)	4,575 / 1,001	health; Black Gill; WSSV; Yellow Head	fused record group; identity unavailable	931	disease
L. vannamei biomass (V2, 10.17632/h8tcn6ykky.2)	5,507 / 5,507	weight; total length	individual animal	170	biomass
Total	11,977 / 7,657	—	—	1,513	—

Per-dataset individual counts, site-grouping information, and the complete set of data columns are provided in Supplementary Section S2. Each collection presents a distinct data-quality concern (Section 4.2). ShrimpDiseaseImageBD contains byte-identical annotated copies without timestamps, whereas augmented derivatives in TigerShrimpBD are not distinguished from original images, and health labels are associated with capture sessions. The

biomass collection contains repeated observations of a relatively small number of animals: 22 animals in DB1, 58 in DB2, 50 in DB3, 20 in DB4, and 20 in DB5. Between 15 and 80 views were recorded per animal, with a median of 30.

Representative images from the disease categories in ShrimpDiseaseImageBD and TigerShrimpBD, together with examples from biomass databases DB1–DB5, are presented in Figure 2. For each stratum, the upper-median file-size image was selected, with ties resolved by record identifier. Images are displayed at their native aspect ratios and pixel dimensions, without illumination enhancement. Full-size examples are provided in Supplementary Section S9. Visible differences in background, illumination, and pose illustrate acquisition variation, although no causal interpretation of these differences is implied.

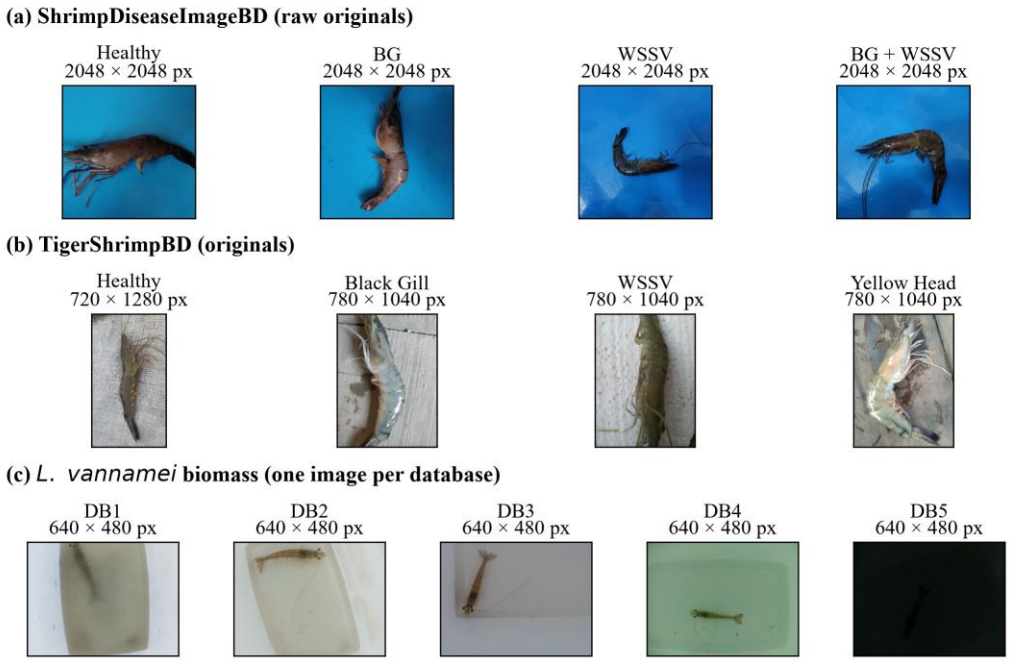


Figure 2. Representative raw images from the three collections

3.2. Problem Formulation and Masked Supervision

A shared encoder f maps an image $\mathbf{x}$ to a feature $\mathbf{z} = f(\mathbf{x})$, which is subsequently processed by task-specific heads to produce the outputs described in Section 3.4. A common label space is used for the two disease datasets, consisting of a health target, with the diseased class treated as positive, and three disease attributes. However, Yellow Head is not annotated in ShrimpDiseaseImageBD. Therefore, the corresponding target is masked for every record from this dataset rather than being treated as a negative label. Similarly, cephalothorax length is unavailable in DB1 and is consequently masked for records from that database.

Each task loss is calculated only over entries with available supervision. A conditional masking operation is applied to exclude unavailable targets from the loss computation, thereby preventing non-finite values associated with masked entries from contributing to the model gradients.

3.3. Data Audit and Leakage-Safe Grouping

A 22-check audit gates training: it verifies files, dimensions, label-mask integrity, measurement joins and filename parsability, measures the relations below, and blocks training if a leakage-critical check fails (Section 4.2).

The grouping key is the connected components of the union of four relations: exact duplicates sharing a SHA-256 content hash; published augmentation families, where an augmented file names its source original, or, in ShrimpDiseaseImageBD, a raw image names its annotated copy; perceptual near-duplicates at Hamming distance at most six, blocked by dataset and class for the disease sets or by dataset and individual for the biomass set; and every image of one identified animal. Cross-dataset near-duplicates are reported but never fused, so the two datasets stay separate.

The fourth relation applies wherever an individual is identifiable; only ShrimpDiseaseImageBD and the biomass collection identify one (Supplementary Section S1). This key keeps all identifiable views of the same animal in one partition.

All splits are built over this key: disease protocols use a stratified group k-fold whose stratum is dataset and class, with validation carved from the training folds by a second stratified group split, and the cross-dataset protocols train on one dataset's pool and test on the other's; biomass protocols hold out a database, a site, or grouped folds within a database. A split is accepted, and its split-membership hash written, only if an automated check finds no group identifier in more than one partition, under the protocol's own key and an independent re-derivation from the manifest. The naive control arm alone skips this fusion, splitting files by a stratified draw over class, so its test partition is group-leaked (Section 3.6).

3.4. Shared-Encoder Architecture and Task Heads

A shared visual encoder is connected to task-specific heads, together with a domain head assigned zero weight in the main configuration and used only for ablation. The encoder output dimension, d, is determined through a forward pass during model initialization because the library-provided dimension is inaccurate for MobileNetV3-Large. Each task head comprises a two-layer perceptron with dimensions d → h →k, where d denotes the encoder output dimension, h the hidden-layer width, and k the task-specific output dimension: one health logit, three disease attributes, or three biomass targets (log weight, total length, and cephalothorax length). Encoder dimensions and hidden-layer widths are reported in Supplementary Section S10, while model size, inference latency, and export parity are provided in Supplementary Section S8.

ConvNeXt-Tiny [28] and the lightweight MobileNetV3-Large [29] are initialized with ImageNet-pretrained weights. All layers are trainable from the first epoch, with the encoder, task heads, and task-weighting parameters included in a single optimizer parameter group. No layers are frozen during training. ResNet-50 [30], EfficientNet-B0 [31] and Swin-T [32] are trained identically (Section 4.5).

3.5. Objectives, Task Weighting and Aggregation

The disease objective, Eq. (1), adds masked attribute and consistency terms to a cross-entropy health loss:

$$\mathcal{L}_{\text{disease}} = \mathcal{L}_{\text{health}} + \lambda_{\text{attr}} \cdot \mathcal{L}_{\text{attr}} + \lambda_{\text{cons}} \cdot \mathcal{L}_{\text{cons}} \tag{1}$$

The attribute loss is masked binary cross-entropy; focal [33] and asymmetric [34] variants are ablated. The consistency term of Eq. (2) penalises an output confident the animal is both healthy and diseased, over supervised attributes only:

$$\mathcal{L}_{\text{cons}} = \frac{1}{B}\sum_{i=1}^{B} \sigma(-h_i) \cdot \max_{k:\, m_{i,k}=1} \sigma(a_{i,k}) \tag{2}$$

B is the batch size, σ the logistic function, h_i the health logit of image i (diseased positive), $a_{i,k}$ the logit of attribute k and $m_{i,k}$ its supervision mask; the maximum is taken as zero when no attribute of an image is supervised. The attribute and consistency weights in Eq. (1) are 1.0 and 0.1, respectively (Supplementary Section S4). Biomass uses a masked Huber loss; task losses combine with unit or homoscedastic uncertainty weights [20]. Per-individual weight is the median of that animal's view predictions. Domain-generalization variants add source-balanced sampling with MixStyle [21] or CORAL [22]; multi-view runs pool views by masked mean or gated attention [35].

3.6. Evaluation Protocols

Table 3 summarizes the evaluation protocols. Disease partitions were constructed with five folds; only folds 0–2 were executed, each at three seeds, because of computational limits. Thus the disease rows report three evaluated folds from a five-fold partition, rather than a complete three-fold cross-validation. The records and groups columns describe the pool included in each constructed partition; folds × seeds counts completed runs. Group-aware disease arms test original images only. The augmentation-enabled arm adds eligible published augmentations to training, whereas the naive arm also permits augmentations in validation

and test. The naive pool contains 5,724 records; the augmentation-enabled group-aware pool contains 4,589–4,601 per fold because augmentations linked to held-out originals are excluded. All other protocols enforce group-based splitting. The metrics and leakage controls are detailed in Supplementary Section S2. Cross-dataset protocols use the shared Healthy, Black Gill, and WSSV label space.

The naive control arm, defined in Section 3.3, is the only protocol that does not enforce group-based separation. Consequently, its test partition does not consist of independent image observations: depending on the fold, 96.4–96.9% of test images belong to groups that also appear in the corresponding training or validation partition, while 61.7–64.2% are published augmented copies rather than original photographs. This arm is intended to represent a commonly used image-level evaluation setting rather than to reproduce a specific published partition, as partition files are not provided by the comparator studies.

The five biomass databases correspond to three physical sites: DB1 represents one farm, DB2 and DB3 represent two sampling cycles at another farm, and DB4 and DB5 represent one academic pond [11]. Leave-one-database-out (LODO) evaluation holds out a sampling occasion within the available sites, whereas leave-one-site-out (LOSO) evaluation holds out an entire physical location. We therefore use LOSO to estimate generalization to an unseen collection site, consistent with the rationale for geographically disjoint validation in another application [36].

Table 3. Evaluation protocols

Protocol	Held-out unit	Records / groups	Folds run × seeds	Role
disease_cv	fused group	2,150 / 1,293	3 × 3	primary within-dataset
disease_cv_withaug	fused group	4,589–4,601 / 1,343	3 × 3	augmentation-enabled arm
disease_cv_naive_withaug	image (control arm)	5,724 / 5,724	3 × 3	naive control arm
disease_cv_sdibd	fused group	1,149 / 412	3 × 3	single-dataset arm
disease_cv_tiger	fused group	1,001 / 881	3 × 3	single-dataset arm
cross_sd2tg	dataset	2,100 / 1,243	1 × 3	cross-dataset test
cross_tg2sd	dataset	1,930 / 1,212	1 × 3	cross-dataset test
biomass_lodo	database	5,507 / 170	5 × 3	biomass, new database
biomass_loso	site	5,507 / 170	3 × 3	biomass, new site
biomass_within_db	individual animal	5,507 / 170	5 × 3	biomass, same database

3.7. Training Configuration

Images were resized to 320 pixels using letterboxing. ConvNeXt-Tiny was used as the primary backbone and optimized with AdamW [37], with a learning rate of 0.0003, a batch size of 12, and 12 training epochs. Data augmentation included horizontal flipping, affine transformations, and photometric adjustments. Three random seeds (42, 2026, and 3407) were consistently applied across all configurations in the experimental grid. All experiments were conducted on a single NVIDIA GeForce RTX 3070 Ti Laptop GPU using Python 3.13.5, PyTorch 2.13.0, and timm 1.0.28.

Except for the multi-task comparisons described in Section 4.4, the reported experiments were conducted using single-task configurations, either for disease recognition or biomass estimation. Additional training configurations are provided in Supplementary Section S4. For reproducibility, each run records content fingerprints of the data manifest and the corresponding data partitions, allowing individual results to be traced to their respective data splits (Supplementary Section S2).

3.8. Evaluation Units and Uncertainty

Disease predictions and performance scores are calculated at the image level using a 0.50 decision threshold, with balanced accuracy as the primary metric. However, data splitting and statistical inference are performed using fused groups of related records and, where available, identified animal-level groups. Individual identifiers are unavailable in TigerShrimpBD.

For biomass estimation, predictions from multiple views are aggregated using the median for each animal, and mean absolute error (MAE) is subsequently calculated at the individual-animal level for each seed. Uncertainty intervals are estimated using 2,000-replicate cluster bootstrap procedures, with disease fused groups or individual biomass animals serving as the resampling units. The resulting 95% confidence intervals represent sampling uncertainty. Detailed results in Supplementary Sections S11 and S12 report the three-seed mean and standard deviation alongside these intervals.

The area under the receiver operating characteristic curve (AUROC) and expected calibration error (ECE) [38] are calculated separately for each seed, after concatenating its evaluated test folds, and then averaged across seeds. Health balanced accuracy at the fixed 0.50 threshold is also calculated from each seed's fold-pooled predictions. Because every seed tests the same records, its seed mean equals the balanced accuracy of the predictions concatenated over all seeds; it is not an unweighted mean of per-fold scores. Multiple comparisons are adjusted using the Holm procedure [39]. Within each biomass protocol, absolute errors are averaged over individuals and then over seeds; database and site folds are weighted by their numbers of tested animals. Different biomass protocols are reported separately. Per-attribute decision thresholds are determined exclusively from validation partitions, and no test partition is used during model selection (Supplementary Section S4).

4. Results and Discussion

4.1. Conventional Within-Dataset Evaluation

A conventional evaluation returns high numbers on this corpus. The naive image-level arm reaches a health balanced accuracy of 0.976 [0.967, 0.983]. Under the grouping of Section 3.3 the pooled arm reaches 0.916 [0.896, 0.934], ShrimpDiseaseImageBD alone 0.812 [0.769, 0.852] and TigerShrimpBD 0.998 [0.995, 1.000], the last a bound on session separability rather than disease recognition (Section 4.2.). A pooled and a single-dataset configuration partition different populations, so a number from one must never be read against a number from the other. The five arms in full are in Supplementary Section S11; Sections 4.2 and 4.3 report what the collections contain and what the protocol alone is worth.

4.2. Data Audit: What the Collections Contain

Of the audit's 22 checks, 16 pass, 5 return warnings, and 1 returns an informational failure. No leakage-critical check fails, so the gate reads CLEAR. All 35 constructed group-aware protocol-fold combinations are disjoint: 20 from four disease protocols with five folds each, 2 cross-dataset folds, and 13 biomass folds (5 LODO, 3 LOSO, and 5 within-database). Of these, 27 combinations were executed; the deliberately leaky image-level control is excluded from both counts. Section 3.3's four relations collapse 11,977 records to 1,513 groups: exact duplicate, 11,230; published augmentation family, 8,732; perceptual near-duplicate, 3,597; forced individual, 1,513. Supplementary Section S2 gives the full funnel, link counts, and an example of each relation.

A file-level grouping comparator leaves 84.8–92.6% of the ShrimpDiseaseImageBD test images whose animal is identifiable sharing that animal with the same fold's training or validation partition. Animal-level grouping reduces this overlap to 0.0% in every group-aware protocol. The ShrimpDiseaseImageBD pool contains 1,089 file-level groups, which merge into 412 animal-level groups over 416 identified individuals. Supplementary Section S2 records the grouping and partition membership fingerprints.

Every diseased TigerShrimpBD original is 1040 × 780 or 780 × 1040 (578 of 578), and no healthy one is either size (0 of 423), so an exact-size lookup separates the classes at 1.000. Metadata separate the two datasets at 1.000, colour moments at 0.977; on ShrimpDiseaseImageBD colour moments reach 0.743. The full probe grid is in Supplementary Section S3.

A filename-timestamp analysis of TigerShrimpBD recovers 13 capture sessions, 8 diseased-only and 5 healthy-only, accounting for all 1,001 originals. The health label is therefore perfectly predictable from the capture date, a property of collection that no splitting strategy can remove, so within-dataset health accuracy there bounds session separability rather than disease recognition. ShrimpDiseaseImageBD filenames carry no timestamp, so there the confound is neither confirmed nor excluded.

4.3. Controlled Protocol Comparison

Table 4 compares five protocol rungs. Intervals are 95% cluster-bootstrap intervals over fused groups; rung 0 contains non-neural probes, rung 2 uses file-level grouping at one seed, and rung 3 uses animal-level grouping at three seeds. Supplementary Section S11 gives the full table and bar chart. The naive image-level arm includes 5,724 records, with 1,145 test records in each evaluated fold (0–2), or 3,435 distinct test records per seed; 96.4–96.9% of these test images share a group with training or validation. Its architecture and hyperparameters match the augmentation-enabled group-aware arm, but its training records and test composition differ. Balanced accuracy is 0.976 [0.967, 0.983] in the naive arm and 0.934 [0.917, 0.952] in the augmentation-enabled group-aware arm, a difference of −0.041 calculated before rounding (0.9341812065 − 0.9755979510 = −0.0414167445). Of the naive arm's test records, 63% are published augmented copies, scored at 0.999. Its originals alone score 0.947, 0.013 above the group-aware arm, which tests originals only. These original-image subsets still contain different records, so neither difference isolates a causal effect of leakage. SD and TG abbreviate ShrimpDiseaseImageBD and TigerShrimpBD (SDIBD also denotes the former), LODO and LOSO denote leave-one-database-out and leave-one-site-out, and ERM denotes empirical risk minimisation.

The file-level and animal-level arms use the same recipe but different grouping keys, which reassign fold membership. Across evaluated folds 0–2, the animal-level pooled arm tests 431 + 431 + 429 = 1,291 distinct original records from its eligible pool of 2,150; its ShrimpDiseaseImageBD-only arm tests 230 + 230 + 229 = 689 records. The remaining originals belong to the two unevaluated test folds. Of these evaluated records, 758 in the pooled arm and 408 in the single-dataset arm are shared with the file-level comparator. On those shared records at seed 42, single-dataset balanced accuracy falls from 0.884 to 0.797 (−0.087), while pooled balanced accuracy rises from 0.904 to 0.927 (+0.023). Thus animal-level separation need not lower every score; full-set differences between the arms are not comparisons on identical test records.

Table 4. Protocol ladder

Rung	Protocol	Configuration	Folds × seeds	BA [95% CI]	AUROC	BA on shared records
0	—	metadata probe	—	0.732	0.856	—
0	—	colour-moment probe	—	0.716	0.794	—
0	—	8 × 8 thumbnail probe	—	0.770	0.847	—
0	—	Tiger size-lookup rule, in-sample	—	1.000	—	—
1	disease_cv_naive_withaug	naive image split	3 × 3	0.976 [0.967, 0.983]	0.998	—
2	disease_cv	files, pooled	3 × 1	0.895 [0.879, 0.913]	0.964	0.904
2	disease_cv_sdibd	files, SDIBD	3 × 1	0.873 [0.845, 0.900]	0.948	0.884
3	disease_cv	animals, pooled	3 × 3	0.916 [0.896, 0.934]	0.973	0.927
3	disease_cv_sdibd	animals, SDIBD	3 × 3	0.812 [0.769, 0.852]	0.886	0.797
3	disease_cv_withaug	group-aware on animals, augmented training pool	3 × 3	0.934 [0.917, 0.952]	0.981	—
4	cross_sd2tg	cross-dataset SD → TG	1 × 3	0.494 [0.480, 0.510]	0.459	—
4	cross_tg2sd	cross-dataset TG → SD	1 × 3	0.380 [0.350, 0.410]	0.271	—

Figure 3 separates the internal protocol contrast from transfer. Both panels show health balanced accuracy at the 0.50 threshold for seeds 42, 2026 and 3407: panel (a) uses the same architecture and hyperparameters with different eligible training records, splits, and test composition, and panel (b) shows within- and cross-dataset group-based protocols. Test populations differ, so the differences are not pure causal leakage magnitudes; Table 4 reports the confidence intervals. The 0.976-to-0.934 difference combines grouping and test composition: 63% of the naive test records are published augmented copies. Its originals alone score 0.947, still on different test records. Cross-dataset testing changes the collection and acquisition distribution as well; its degradation is evidence of limited transfer, not an additional quantity of leakage. The shared-record regrouping comparison also moves in opposite directions across the pooled and single-dataset arms, so stricter grouping need not lower every score.

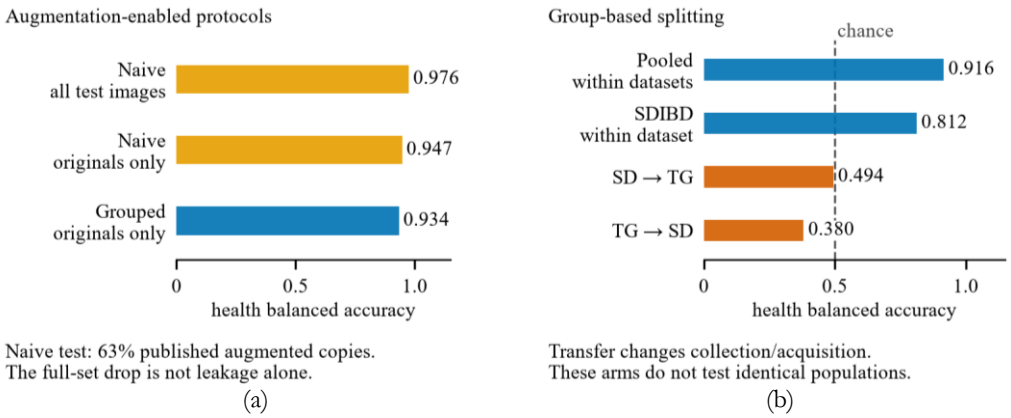


Figure 3. Health balanced accuracy (**a**) across split protocols and (**b**) within and across datasets.

4.4. Controlled Ablation

We test eighteen paired cluster-bootstrap comparisons (loss, sampling, augmentation, task weighting, MixStyle, CORAL, backbone, multi-view pooling and protocol) over three seeds; Table 5 prints all eighteen — the pre-specified ten and the eight exploratory ones, since a table filtered on its own result would invite the reader to mistake a selection for a sweep. In Table 5, $\Delta = A - B$ is the health balanced accuracy for disease (higher is better) and the individual-weight MAE in grams for biomass (lower is better), taken as the mean paired difference over seeds 42, 2026 and 3407; n is the shared record count, or the animal count for biomass. The 95% CI uses 2,000 paired cluster-bootstrap replicates over fused groups or animals, and † marks $p < 0.05$ at or below the bootstrap floor (0.006 for disease, 0.003 for biomass). Holm adjusts within family (twelve disease, six biomass, exploratory rows included); one survives at 0.05: within-database CV against leave-one-DB-out, −3.15 g, Holm $p \leq 0.003$ at the bootstrap floor, a protocol effect, not a component effect. Every disease Holm p is 1.000. Section 4.5 tests MixStyle across datasets.

Thirteen intervals exclude zero without a Holm rejection: each gives the sampling precision of three fixed model pairs with no training-seed component; the directional Holm p asks whether every seed agreed; the dagger follows Holm alone. Eleven of eighteen contrasts change sign across seeds; of those eleven, two have all three per-seed intervals clear of zero and six have two or more significant seeds.

Holm p is a directional intersection–union p: it takes the least favourable of the three seeds' own two-sided bootstrap p-values, so a contrast is only as strong as its weakest seed and saturates at 1 as soon as one seed reverses. The extended table of Supplementary Section S15 prints the same eighteen rows with each seed's own p-value and the envelope of the three seeds' intervals beside this one, the fold × seed signed-rank p, the confirmatory/exploratory tag, and the two configurations each comparison contrasts, which carry a row into the configuration-to-ablation map of Supplementary Section S4. That signed-rank p is the test over paired (fold, seed) cells and is not computable on four of the rows, whose two runs hold out different records; where it is, it cannot fall below $2/2^n$ (0.0039 at $n = 9$ and 0.000061 at $n = 15$), so a value near that limit is a resolution floor rather than a result. The eight exploratory rows — loss variants, sampling, colour-augmentation strength, dataset pooling, multi-view pooling — change nothing: not one reaches 0.05 under Holm.

The robust ablation finding is a protocol effect: within-database evaluation lowers weight MAE by 3.15 g against leave-one-database-out. None of the twelve disease contrasts survives Holm correction, and no individual architectural or training component establishes a stable improvement across the three seeds. The eight exploratory contrasts describe unstable differences, not component gains; pooled intervals do not account for variation from retraining.

Masked supervision (Supplementary Section S5): adding 689 fabricated Yellow-Head negatives across evaluated folds 0–2 of the five-fold partition, scored against 10 Yellow-Head positives in each fold (30 of the 50 available originals), with three seeds' predictions pooled, lowers average precision, the threshold-free pre-registered endpoint, in every fold (0.885→0.694, 0.861→0.627, 0.970→0.904); letterbox padding separates the sources by appearance, making a null uninterpretable. MobileNetV3-Large's +0.036 is a pooled-protocol

difference; the lightweight backbone has no TigerShrimpBD-only run; different partitions must never be ranked across protocols. Configuration-to-ablation map: Supplementary Section S4; tabular tier: Supplementary Section S6.

Table 5. Controlled ablation of eighteen paired comparisons

Comparison (A − B)	Task	n	Δ mean [95% CI]	Holm p
pooled vs single-dataset	disease	364	−0.008 [−0.015, −0.002]	1.000
pooled vs single-dataset	disease	434	+0.025 [−0.004, +0.053]	1.000
masked asymmetric loss vs BCE	disease	1,291	−0.006 [−0.017, +0.005]	1.000
masked focal loss vs BCE	disease	1,291	−0.015 [−0.026, −0.005]	1.000
source-balanced vs ordinary sampling	disease	1,291	−0.022 [−0.033, −0.011]	1.000
with published augmented copies vs originals only	disease	1,291	+0.019 [+0.007, +0.030]	1.000
aggressive colour augmentation vs disease-safe	disease	1,291	−0.004 [−0.014, +0.007]	1.000
multi-task vs single-task disease	disease	1,291	−0.044 [−0.061, −0.027]	1.000
uncertainty weighting vs fixed task weights	disease	1,291	+0.035 [+0.022, +0.049]	1.000
MixStyle vs ERM multi-task	disease	1,291	−0.040 [−0.058, −0.022]	1.000
CORAL vs ERM multi-task	disease	1,291	+0.036 [+0.021, +0.051]	1.000
MobileNetV3-Large vs ConvNeXt-Tiny	disease	1,291	+0.036 [+0.023, +0.049]	1.000
multi-view mean pooling vs single image	biomass	170	+1.66 [+0.67, +2.63]	1.000
attention pooling vs mean pooling	biomass	170	+0.39 [−0.45, +1.28]	1.000
multi-task biomass vs single-task biomass	biomass	130	+1.56 [+1.04, +2.13]	0.656
MobileNetV3-Large vs ConvNeXt-Tiny	biomass	170	−2.10 [−3.09, −1.11]	1.000
within-database CV vs leave-one-DB-out	biomass	170	−3.15 [−4.07, −2.30]	≤0.003 †
leave-one-site-out vs leave-one-DB-out	biomass	170	+2.87 [+2.25, +3.49]	0.125

4.5. Cross-Dataset Testing

Under rung 4's cross-dataset protocol the configuration reported throughout this paper collapses in both directions: training on ShrimpDiseaseImageBD and testing on TigerShrimpBD gives a health balanced accuracy of 0.494 [0.480, 0.510] at an area under the curve of 0.459, and the reverse direction 0.380 [0.350, 0.410] at 0.271. Neither balanced accuracy clears the chance level of 0.50, and only TG → SD lies wholly below it; both areas under the curve are anti-correlated with the label on the three-seed mean, and calibration error rises to 0.393 and 0.574: the transferred models are confidently wrong. MixStyle lifts balanced accuracy to 0.544 and 0.444, outside the untreated intervals in both directions, only SD → TG clearing the 0.50 chance level, and moves the SD → TG ranking from 0.459 to 0.586. Figure 4 shows representative class-specific errors; Supplementary Section S11 gives the full matrices.

The cross-dataset sweep uses ConvNeXt-Tiny, ResNet-50, EfficientNet-B0, and Swin-T, plus a second ConvNeXt-Tiny configuration trained with MixStyle: four encoder families and five configurations (Supplementary Section S11). MobileNetV3-Large belongs to the separate pooled lightweight experiment, not this sweep. The collapse differs by direction. In the TG → SD direction no encoder ranks the target above chance, the area under the curve spanning 0.271–0.390, and no threshold repairs it: the best balanced accuracy any single threshold reaches, chosen with the target labels in hand, is only 0.405–0.500. Four encoder families spanning 4.7 to 28.2 million parameters all rank below chance there, at every seed, which capacity among these encoders does not explain. SD → TG separates by encoder instead: EfficientNet-B0 carries its ranking of 0.751 through to a deployed balanced accuracy of 0.615 [0.595, 0.636], within 0.093 of its own oracle ceiling and at the lowest calibration error in the sweep, 0.219, while ResNet-50 ranks best of all at 0.834 and still scores only 0.553 at its source-selected threshold, with a calibration error of 0.308 against an oracle-threshold 0.765. The SD → TG direction is a property of the encoder and of the threshold; the encoder this paper reports elsewhere is the worst of the five configurations in it. Every row of that sweep is a single fold at seeds 42, 2026 and 3407.

Across datasets the within-dataset ordering does not survive: Black Gill, weakest within the pooled protocol, ranks at 0.620 in SD → TG, while WSSV, the stronger, sees recall fall to 0.080 in the reverse. The per-attribute table is in Supplementary Section S11.

Figure 4 makes the collapse visible in the healthy class: at representative seed 42, 86% of TigerShrimpBD healthy images are called diseased after SD → TG training, and 88% of ShrimpDiseaseImageBD healthy images after TG → SD. The reverse arm also misses 40% of diseased images. These are image counts from one seed, not the three-seed means above, and show the error balance hidden by one accuracy. The matrices use a 0.50 threshold; SDIBD pools folds 0–2, and SD → TG and TG → SD use fold 0. Cells show image counts and true-class-normalised percentages on a shared 0–100% scale. The representative balanced accuracies are 0.789, 0.461 and 0.362, respectively, and differ from the three-seed means; the acquisition-confounded Tiger within-dataset matrix remains in Supplementary Section S11.

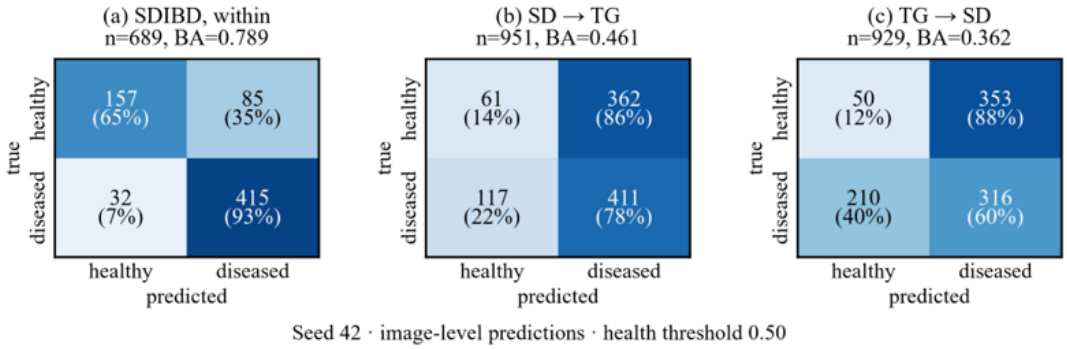


Figure 4. Representative single-task health confusion matrices at seed 42

4.6. Biomass Estimation across Databases and Sites

Table 6 reports every held-out unit: the image-only model beats the train-mean predictor on all 13 on the three-seed mean, 10 at every seed, cutting its error by 3–92% (median 46%); on that mean it beats a database-mean oracle on 0 of 5 databases and 0 of 3 sites (four units at single seeds, by ≤ 0.15 g), and within a database on 2 of 5 by 0.074–0.123 g, inside the intervals (Supplementary Section S12). Over 170 animals the error is 3.18 g within database, 6.33 g with a database held out, 9.20 g with a site held out, the site split costing 2.87 g [2.25, 3.49] more; the five databases are three sites, so the site figure is defensible. Centred on each database's mean, the coefficient of determination is 0.210 within database, −2.493 across them.

Table 6. Individual-level biomass error by held-out unit

Protocol	Held-out unit	Animals	Image MAE (g)	Train-mean MAE (g)	DB-mean oracle MAE (g)	RF on true length MAE (g)
LODO	DB1	22	3.93	7.22	3.83	3.40
LODO	DB2	58	3.87	8.86	3.80	2.20 (2.58)
LODO	DB3	50	3.97	8.41	3.79	2.28 (2.49)
LODO	DB4	20	17.96	18.51	1.06	0.58 (0.54)
LODO	DB5	20	10.35	18.52	1.10	0.61 (0.59)
LOSO	DB1	22	4.06	7.22	3.83	3.40
LOSO	DB2, DB3	108	6.73	14.10	3.79	19.40 (3.87)
LOSO	DB4, DB5	40	18.70	21.70	1.08	6.51 (2.60)
Within-DB	DB1	22	3.76	6.41	3.83	3.02
Within-DB	DB2	58	3.67	6.85	3.80	2.21 (2.42)
Within-DB	DB3	50	3.88	6.39	3.79	2.33 (2.26)
Within-DB	DB4	20	1.29	16.42	1.06	0.60 (0.58)
Within-DB	DB5	20	1.24	16.37	1.10	0.85 (0.84)

In Table 6, LODO holds out a database, LOSO a site and Within-DB individual animals, and rows never pool; each value is the mean over three seeds. Beside the model stand the train-ing-set mean, a held-out database-mean oracle and a random forest given the ground-truth manual length, which is unavailable to image-only inference. The forest uses length and cephalothorax measurement when both sides carry it; parentheses give the length-only refit

where the rule changed the fit. Centred R2 measures variance beyond the database mean and, on one database, equals plain R2. Supplementary Section S12 gives the bootstrap intervals, seed mean ± SD, image R2, centred R2 and allometric error, and Supplementary Section S7 the per-database error.

Figure 5 averages the three seeds' animal-level predictions from the leave-one-database-out runs of Table 6, and neither panel tracks the identity line. Each point is one animal (DB1 22, DB2 58, DB3 50, DB4 20, DB5 20), coloured by database; view medians are computed within each seed, and predictions are then averaged over seeds 42, 2026 and 3407. The dashed line is y = x, and Supplementary Section S12 decomposes the error. On DB1, DB2 and DB3 the cloud is horizontal: observed weights span 1–41 g and the predictions do not follow them, which is consistent with a coefficient of determination between −0.044 and 0.046. On DB4 and DB5, whose combined site spans 6–12 g (DB4: 7–12 g; DB5: 6–11 g), the points leave the line altogether: the error there, 17.96 and 10.35 g, is larger than the entire observed weight range of those databases (5 g in both), so the cloud sits further from the line than the animals sit from each other. What the model recovers is a pool's level, not an animal's size: hence the database-mean oracle, not the training-set mean, is the reference this subsection reads against.

With the same animal weighting as the image-model errors, the forest given manual length measurements has MAE 2.00 g in leave-one-database-out evaluation and 14.30 g in leave-one-site-out evaluation. Three limitations explain this sensitivity: the forest cannot extrapolate outside its training range (for DB2 and DB3, training animals span 6–12 g while test animals span 1–39 g); the feature rule excludes training animals without a cephalothorax measurement, which is unavailable for DB1; and the length–weight relationship differs between sites, predicting 8.8–9.0 g at DB4 and DB5 versus 11.3–15.6 g elsewhere at the same reference length. A length-only refit that retains animals without cephalothorax measurements has animal-weighted LOSO MAE 3.51 g. Given manual size, regression is largely recoverable within a site (mean R^2 0.630 over the five held-out databases, 0.329–0.819 per database). These image-only models do not recover absolute size across held-out databases. Manual measurements are unavailable to image-only inference. Adding image features to length changes animal-weighted LODO MAE by +0.2599 g.

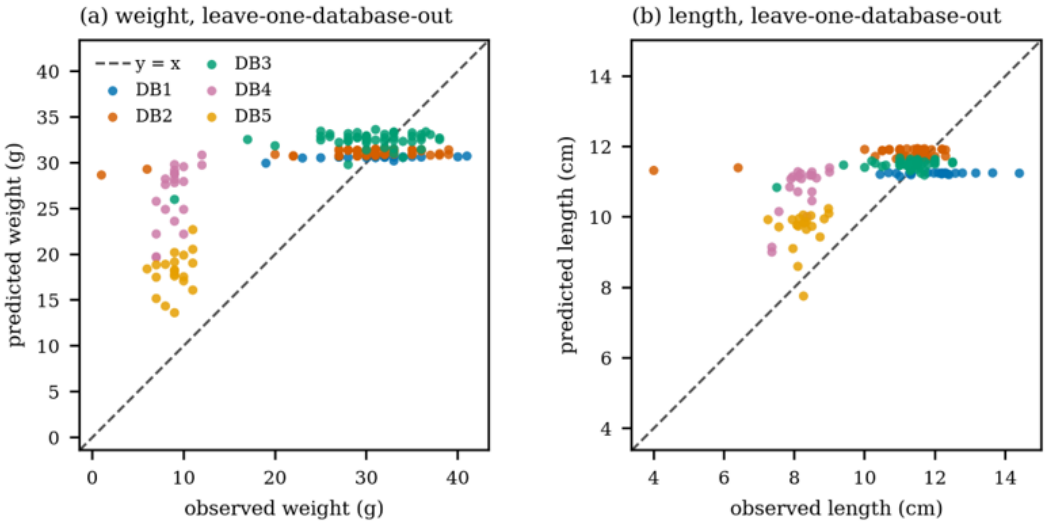


Figure 5. Predicted against observed weight (**a**) and length (**b**) with a database held out

Acquisition geometry is unrecorded: the provider logs two manual lengths, weight and image; all 5,507 images are 640 × 480 pixels. No scale reference is documented; none appears in the five frames we opened, one per database: one animal from above in a shallow tray of water, no ruler, marker or object of known size. This is a limitation of the recorded data. The results are consistent with unresolved absolute scale, but no controlled change of camera distance or scale calibration was tested, so the acquisition analysis does not establish a causal scale effect.

4.7. Acquisition Evidence and Interpretation

The acquisition-session confound of Section 4.2. fixes the TigerShrimpBD health label independently of the animal, consistent with the collapse in Section 4.5; Section 4.2's audit adds an exact-size lookup and dataset-naming metadata. A further distinction is acquisition: no two collections share a dominant pixel size, and inside the biomass collection a linear probe on an 8 × 8 grey thumbnail recovers an image's database at 0.993 balanced accuracy over folds grouped by individual, against 0.996 from a colour thumbnail and 0.997 from six colour moments in its within-database probes (Section 4.2). Conditions differ substantially between and within collections; the proxies do not separate the geometric component from the photometric: camera, working distance and illumination are unrecorded.

Identity annotation exists in only two collections — 416 animals in ShrimpDiseaseImageBD, 170 in the biomass one, none in TigerShrimpBD. Disease cues largely do not survive a change of collection: only Black Gill clears chance in either direction; the rest sit near it (Section 4.5), and two of the three signs are colour, which the database probe above shows is collection-specific. The biomass results are consistent with unresolved absolute scale (Section 4.6). Supplementary Section S13 gives the acquisition statistics and both models' Grad-CAM maps [40].

4.8. Threats to Validity

The acquisition-session confound of Section 4.2. bounds the TigerShrimpBD result: a balanced accuracy of 0.998 there measures session separability, not disease recognition, and in ShrimpDiseaseImageBD it is untestable rather than absent. Attribute support is thin: TigerShrimpBD holds 39 Black Gill and 50 Yellow Head originals. Biomass rests on 170 animals, with 5 of 5 databases and 3 of 3 sites held out in turn (Section 4.6), at the substituted collection version of Section 3.1. The disease grid is small — evaluated folds 0–2 of five-fold partitions at three seeds, with one fold at three seeds across datasets and in the backbone sweep — and a Wilcoxon test over n paired folds cannot reach a two-sided p below $2/2^n$ (0.25 at three pairs, 0.0039 at nine), so the fold-level tests are descriptive.

Saved-run provenance flags are incomplete, and reproducibility is approximate rather than bitwise. Two seed-42 TG → SD fits on identical partitions give balanced accuracies of 0.355 and 0.362 and areas under the curve of 0.303 and 0.258. Supplementary Section S2 records the available provenance and membership fingerprints. These experiments support no clinical or farm-deployment claim.

5. Comparison with Reported Results

Table 7 sets ours beside published numbers in two blocks answering different empirical questions: whether a model separates one collection's classes, and whether that separation survives a leakage-aware partition. BA is the health balanced accuracy at 0.50 (Section 3.8), and the biomass contrast uses the same 170 animals. Values are quoted as published; a value taken from an abstract is labelled, and an unverifiable one is marked Inaccessible. Supplementary Section S14 gives the full protocols, grouping and intervals, and Supplementary Section S11 gives the class-count discrepancy of the first study and its omitted co-infection class. Matched on protocol, rung 1 — those studies' random image-level split — reaches 0.976 [0.967, 0.983] health-head balanced accuracy, against four-class 93% ± 0.059 accuracy on these collections [4] and 100% accuracy on a private SEAFDEC/AQD-and-Roboflow set [3]; rung 3, group-aware, gives 0.934 [0.917, 0.952].

Neither block may be read down its result column. In the first, each published figure is that study's own plain accuracy over its own class set — four classes for [4], a White Spot Syndrome Virus result for [3], slide-grouped histopathology patches for [13] — while both of ours are the balanced accuracy of the health head at the fixed 0.50 operating point of Section 3.8, a two-class target. The rows put each protocol's result on the record beside the others; the magnitudes are not comparable, and the contrast the block supports is the one between rung 1 and rung 3, which use the same architecture and hyperparameters but differ in eligible training records, split, and test composition.

One row of the second block is a published cross-dataset result rather than one of ours. [4] report, in the limitations of the same paper whose five-fold accuracy the first block quotes, a model trained and validated on an 8:2 split of ShrimpDiseaseImageBD and tested on TigerShrimpBD over the three classes the two collections share, Yellow Head excluded. The

direction and the sign of the degradation match our ShrimpDiseaseImageBD-to-TigerShrimpBD arm; the magnitudes do not. Theirs is plain accuracy over three classes, ours the balanced accuracy of the health head; the class counts they report for TigerShrimpBD are the archive's augmented copies, while our matching arm scores the 951 originals that make up its own held-out target; and their source model is fitted on one 8:2 split with no individual identifiers, not on the group-aware training pool used here.

The fall comes with the protocol: our configuration transfers at 0.494 into TigerShrimpBD, 0.380 into ShrimpDiseaseImageBD, and no encoder in the sweep ranks the latter above chance. Neither shows earlier work to be wrong. TigerShrimpBD ships 1,001 originals and 3,574 augmented copies in per-class original/ and augmented/ subfolders, so a random split over all files leaks, while the layout allows a split over the originals alone — an inference about protocols, not about a study.

Table 7. Our results beside published results

Study	Protocol and grouping	Result
	Within-dataset, conventional protocol	
Sharma et al. [4], same datasets	Stratified 5-fold random image split; individual grouping not reported	93% ± 0.059 accuracy
Eder et al. [3], WSSV	5-fold over originals and augmented copies, on a private SEAFDEC/AQD-and-Roboflow set	100% accuracy, as its abstract reports
Büyükarıkan [12]	5-fold; split-independence details not independently verified	97.3% ± 0.4% accuracy, abstract only
Aung et al. [13]	Slide-grouped patches; histopathology, not gross morphology	99.74% accuracy
This work, rung 1	Image-level split; folds 0–2 of five; 3,435 test records per seed; 96.7% share a group with train/validation	0.976 [0.967, 0.983] BA
This work, rung 3	Group-aware over animals; eligible training augmentations; originals tested; duplicates fused	0.934 [0.917, 0.952] BA
	Leakage-aware and cross-dataset protocol	
Sharma et al. [4], cross-dataset	8:2 ShrimpDiseaseImageBD split, tested on TigerShrimpBD	0.7416 target accuracy, 0.138 gap
This work, both directions	Whole dataset held out; source-side threshold	Into TigerShrimpBD BA 0.494, AUROC 0.459; into ShrimpDiseaseImageBD 0.380, 0.271
This work, four encoders + MixStyle	Same protocol; the encoder or its normalisation changes	Below chance into ShrimpDiseaseImageBD, AUROC 0.271–0.390; TigerShrimpBD 0.459–0.834
Chirdchoo et al. [15], clay pond	One feed-tray rig, whose geometry supplies the scale	R2 = 94.50%
This work, biomass	Leave-one-site-out, 3 sites, 170 animals	MAE 9.20 g by site, 6.33 by database, 3.18 within
This work, forest on length	Same folds; no image, manual measurements only	Animal-weighted MAE 14.30 g; length only 3.51 g; allometric law 5.60 g

6. Conclusions

This study demonstrates that the reliability of image-based shrimp analysis on assembled public datasets depends not only on model selection but also, and often more substantially, on evaluation protocols and data acquisition conditions. The leakage-aware experiments showed that grouping related images before partitioning can substantially change the estimated disease-recognition performance compared with conventional image-level evaluation. However, controlling data overlap does not eliminate acquisition-related confounding. The findings, particularly the near-perfect within-dataset performance observed for TigerShrimpBD, indicate that models may exploit collection-specific characteristics rather than

disease-related visual features. These results highlight the need to interpret classification performance alongside explicit audits of data relationships and acquisition conditions.

The cross-dataset experiments further demonstrated that disease-recognition performance does not transfer reliably between independently collected datasets. Transferability was strongly dependent on the direction of transfer and the selected encoder, indicating that performance within one dataset cannot be taken as evidence of generalisation to another collection. Although some configurations improved cross-dataset performance, the results remained inconsistent across transfer directions. This suggests that differences in acquisition conditions and dataset characteristics remain important barriers to reliable deployment. For biomass estimation, monocular images provided useful information about broad population-level weight patterns, but did not support consistently reliable individual-level estimation under database- and site-held-out evaluation. The increase in mean absolute error from 3.18 g in within-database evaluation to 6.33 g in database-held-out evaluation and 9.20 g in site-held-out evaluation further illustrates the difficulty of transferring image-based weight estimates across collection conditions. Without a calibrated scale reference or other geometric information, absolute body size remains insufficiently constrained by a single image.

Several limitations should be considered when interpreting these findings. The analysis was restricted to three public datasets with different acquisition conditions and annotation schemes, while health status was not systematically represented across capture sessions. Furthermore, the absence of controlled camera geometry and scale calibration limited the assessment of absolute body measurements for biomass estimation. Future work should therefore use more systematically designed datasets that account for variations in health status and acquisition conditions, and investigate calibrated scale references, controlled imaging geometry, and multi-view or depth information. Overall, performance on assembled public shrimp image collections should be interpreted in light of potential data leakage, acquisition-related confounding, cross-dataset transferability, and the physical limitations of image-based measurement before generalisation to practical aquaculture applications can be established.

Author Contributions: Conceptualization, B.J.; methodology, B.J.; software, B.J.; investigation, B.J.; formal analysis, B.J.; writing—original draft preparation, B.J.; supervision, M.H.W.; validation, M.H.W.; writing—review and editing, M.H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All three datasets are public under a Creative Commons Attribution 4.0 licence: ShrimpDiseaseImageBD, doi: 10.17632/jhrtdj9txm.3; TigerShrimpBD, doi: 10.17632/9dj4sk5d55.1; the Litopenaeus vannamei biomass dataset, doi: 10.17632/h8tcn6ykky.2. Code, split manifests, saved predictions and numerical supplements are available at https://github.com/SeedFlora/udang; the numerical supplement is at https://github.com/SeedFlora/udang/tree/main/supplement. The complete Supplementary Material (Sections S1–S15) is publicly available in PDF and editable DOCX at https://github.com/SeedFlora/udang/tree/main/supplement/full.

Acknowledgments: The authors used Claude (Anthropic) for brainstorming and language editing, including paraphrasing. The authors reviewed and verified all content and take full responsibility for it.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] M. Iftehimul, N. A. Hasan, D. Bass, A. Bashar, M. M. Haque, and M. Santi, "Combating White Spot Syndrome Virus (WSSV) in Global Shrimp Farming: Unraveling Its Biology, Pathology, and Control Strategies," *Viruses*, vol. 17, no. 11, p. 1463, Oct. 2025, doi: 10.3390/v17111463.

[2] N. Jesintha, N. Jayakumar, and K. Madhavi, "Length-weight relationship and Fulton's condition factor of two penaeid shrimp species, Fenneropenaeus indicus and Penaeus monodon of Pulicat Lake, southeast coast of India (Decapoda, Dendrobranchiata)," *Crustaceana*, vol. 98, no. 3, pp. 283–293, Apr. 2025, doi: 10.1163/15685403-bja10444.

[3] M. S. Eder, K. C. Ampolitod, N. H. Bolao, R. J. A. Pajal, and E. M. D. Racho, "SHRIMPAI: A Mobile Application for the Early Detection of White Spot Syndrome Virus in Shrimp Using Convolutional Neural Network," *Mindanao J. Sci. Technol.*, vol. 22, no. 2, pp. 196–220, Dec. 2024, doi: 10.61310/mjst.v22i2.2189.
[4] S. Sharma *et al.*, "A lightweight deep learning architecture for automatic shrimp disease classification," *Sci. Rep.*, vol. 16, no. 1, p. 13837, Mar. 2026, doi: 10.1038/s41598-026-44195-z.
[5] Y. Ektefaie, A. Shen, D. Bykova, M. G. Marin, M. Zitnik, and M. Farhat, "Evaluating generalizability of artificial intelligence models for molecular datasets," *Nat. Mach. Intell.*, vol. 6, no. 12, pp. 1512–1524, Dec. 2024, doi: 10.1038/s42256-024-00931-6.
[6] M. Xu, J.-E. Park, J. Lee, J. Yang, and S. Yoon, "Plant disease recognition datasets in the age of deep learning: challenges and opportunities," *Front. Plant Sci.*, vol. 15, p. 1452551, Sep. 2024, doi: 10.3389/fpls.2024.1452551.
[7] B. G. Hill, F. L. Koback, and P. L. Schilling, "The risk of shortcutting in deep learning algorithms for medical imaging research," *Sci. Rep.*, vol. 14, no. 1, p. 29224, Nov. 2024, doi: 10.1038/s41598-024-79838-6.
[8] L. D. Tuan, L. N. An, and D. D. Thai, "Beyond Reported Accuracy: A Verifiability-Focused, Multi-Axis Review of Content and Context Based Fake News Detection," *J. Comput. Theor. Appl.*, vol. 4, no. 2, pp. 475–501, Sep. 2026, doi: 10.62411/jcta.17476.
[9] M. M. Islam *et al.*, "ShrimpDiseaseBD: An image dataset for detecting shrimp diseases in the aquaculture sector of Bangladesh," *Data Br.*, vol. 60, p. 111553, Jun. 2025, doi: 10.1016/j.dib.2025.111553.
[10] S. I. Ahmed and D. M. Farid, "TigerShrimpBD: A Tiger Shrimp Image Dataset," 2025, *Mendeley Data*. doi: 10.17632/9dj4sk5d55.1.
[11] F. J. Ramírez-Coronel, E. Esquer-Miranda, O. M. Rodríguez-Elias, P. García-Hinostro, and G. C. Parra-Salazar, "A Litopenaeus vannamei shrimp dataset for artificial intelligence-based biomass estimation and organism detection algorithms," *Data Br.*, vol. 57, p. 110964, Dec. 2024, doi: 10.1016/j.dib.2024.110964.
[12] B. Büyükarıkan, "Robust shrimp disease detection using multi-model convolutional neural networks-based ensemble strategies," *Aquac. Eng.*, vol. 111, p. 102616, Oct. 2025, doi: 10.1016/j.aquaeng.2025.102616.
[13] T. Aung, R. Vanichviriyakit, K. Chayantrakom, S. Amornsamankul, and P. Huabsomboon, "CNN-Based Identification of Pathogens of Concern in Shrimp," *Animals*, vol. 15, no. 21, p. 3194, Nov. 2025, doi: 10.3390/ani15213194.
[14] R. K. Rachman, D. R. I. M. Setiadi, A. Susanto, K. Nugroho, and H. M. M. Islam, "Enhanced Vision Transformer and Transfer Learning Approach to Improve Rice Disease Recognition," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 446–460, Apr. 2024, doi: 10.62411/jcta.10459.
[15] N. Chirdchoo, S. Mukviboonchai, and W. Cheunta, "A deep learning model for estimating body weight of live pacific white shrimp in a clay pond shrimp aquaculture," *Intell. Syst. with Appl.*, vol. 24, p. 200434, Dec. 2024, doi: 10.1016/j.iswa.2024.200434.
[16] N. A. Ubina, S.-C. Cheng, C.-C. Chang, S.-Y. Cai, H.-Y. Lan, and H.-Y. Lu, "Intelligent Underwater Stereo Camera Design for Fish Metric Estimation Using Reliable Object Matching," *IEEE Access*, vol. 10, pp. 74605–74619, 2022, doi: 10.1109/ACCESS.2022.3185753.
[17] A. Setiawan, H. Hadiyanto, and C. E. Widodo, "Shrimp Body Weight Estimation in Aquaculture Ponds Using Morphometric Features Based on Underwater Image Analysis and Machine Learning Approach," *Rev. d'Intelligence Artif.*, vol. 36, no. 6, pp. 905–912, Dec. 2022, doi: 10.18280/ria.360611.
[18] J. Zhang, S. Zhang, X. Liu, X. Feng, Y. Liao, and T. Pan, "Monocular fish body length estimation via brightness attenuation and fish eye scale structural constraints: application in industrial recirculating aquaculture systems," *Aquac. Int.*, vol. 34, no. 6, p. 210, Jul. 2026, doi: 10.1007/s10499-026-02610-2.
[19] T. Durand, N. Mehrasa, and G. Mori, "Learning a Deep ConvNet for Multi-Label Classification With Partial Labels," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 647–657. doi: 10.1109/CVPR.2019.00074.
[20]A. Kendall, Y. Gal, and R. Cipolla, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 7482–7491. doi: 10.1109/CVPR.2018.00781.
[21] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain Generalization with MixStyle," in International Conference on Learning Representations (ICLR), 2021. [Online]. Available: https://openreview.net/forum?id=6xHJ37MVxxp
[22] B. Sun and K. Saenko, "Deep CORAL: Correlation Alignment for Deep Domain Adaptation," in *Computer Vision -- ECCV 2016 Workshops*, in Lecture Notes in Computer Science. , Cham, Switzerland: Springer, 2016, pp. 443–450. doi: 10.1007/978-3-319-49409-8_35.
[23] A. Ahmad, A. El Gamal, and D. Saraswat, "Toward Generalization of Deep Learning-Based Plant Disease Identification Under Controlled and Field Conditions," *IEEE Access*, vol. 11, pp. 9042–9057, 2023, doi: 10.1109/ACCESS.2023.3240100.
[24] A. Fitzgerald, C. C. Ioannou, S. Consuegra, A. Dowsey, and C. Garcia de Leaniz, "Machine Vision Applications for Welfare Monitoring in Aquaculture: Challenges and Opportunities," *Aquac. Fish Fish.*, vol. 5, no. 1, p. e70036, Feb. 2025, doi: 10.1002/aff2.70036.
[25] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.
[26] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nat. Mach. Intell.*, vol. 3, no. 7, pp. 610–619, May 2021, doi: 10.1038/s42256-021-00338-7.
[27] M. Roberts *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nat. Mach. Intell.*, vol. 3, no. 3, pp. 199–217, Mar. 2021, doi: 10.1038/s42256-021-00307-0.
[28] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.
[29] A. Howard *et al.*, "Searching for MobileNetV3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea: IEEE, Oct. 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[31] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proceedings of the 36th International Conference on Machine Learning, PMLR, vol. 97, 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[32] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy: IEEE, Oct. 2017, pp. 2999–3007. doi: 10.1109/ICCV.2017.324.

[34] T. Ridnik *et al.*, "Asymmetric Loss For Multi-Label Classification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 82–91. doi: 10.1109/ICCV48922.2021.00015.

[35] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," in Proceedings of the 35th International Conference on Machine Learning, PMLR, vol. 80, 2018, pp. 2127–2136. [Online]. Available: https://proceedings.mlr.press/v80/ilse18a.html

[36] M. R. Maulana, P. Sudira, P. Priyanto, Y. A. Fadlullah, D. Nurdiana, and A. H. Bin Sofhi @ Subhi, "Explainable Machine Learning for Predicting Indonesian Vocational School Accreditation: Geographic Validation, Probability Calibration, and Subgroup Auditing," *J. Comput. Theor. Appl.*, vol. 4, no. 2, pp. 502–516, Sep. 2026, doi: 10.62411/jcta.17721.

[37] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Jan. 04, 2019. doi: 10.48550/arXiv.1711.05101.

[38] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proceedings of the 34th International Conference on Machine Learning, PMLR, vol. 70, 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[39] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: https://www.jstor.org/stable/4615733

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.