Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliability-Guided Fourier Alignment and Grouped Multiscale Attention for Sheep Behavior Detection

Yixiang ZENG
Northwest A&F University
Hongjie LIU
Northwest A&F University
HU Jian
Xi'an Xiyangyang Biotechnology Co., Ltd.
WANG Meili
wml@nwsuaf.edu.cn

Northwest A&F University

Research Article

Keywords: sheep behavior detection, YOLO Master, Fourier angle alignment, efficient multiscale attention, image counting

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11704779/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Reliability-Guided Fourier Alignment and Grouped Multiscale Attention for Sheep Behavior Detection

ZENG Yixiang[1], LIU Hongjie[1], HU Jian[2], WANG Meili[1*]

1. College of Information Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China

2. Xi'an Xiyangyang Biotechnology Co., Ltd., Xi'an 710089, China[†]

Abstract: Sheep behavior detection from images is challenging when animals overlap and visually similar behaviors must be distinguished. This study evaluates a P3 feature-fusion adaptation of YOLO-Master that combines Fourier angle alignment (FAA) with efficient multiscale attention (EMA) for four-class detection. Across three paired seeds, the FAA+EMA configuration achieved a mean test AP of 54.61% (SD 3.36), compared with 52.46% (SD 1.37) for YOLO-Master, with 3,746 additional parameters, a 0.133% increase. All three paired AP differences were positive, at 0.48, 5.61, and 0.35 percentage points. However, the mean gain of 2.14 percentage points had a 95% confidence interval from −5.30 to 9.59 percentage points. Because the interval includes zero and the mean was strongly influenced by one seed, the result does not establish a statistically significant mean AP improvement. The largest class-specific gains were observed for Resting and Running, at 4.47 and 4.85 percentage points, respectively. Feeding increased by 0.43 points, whereas Walking decreased by 1.18 points. FAA alone reduced image-level count MAE from 0.549 to 0.401, while the joint configuration had an MAE of 0.574. The projection control reached 54.45% AP, only 0.16 points below FAA+EMA. YOLO26n achieved higher AP with fewer parameters and lower latency than the joint configuration. These findings support a parameter-light, directionally consistent, class-specific and counting-relevant adaptation within YOLO-Master. They also show that the preferred configuration depends on whether detection AP or image-level counting is the primary objective.

Key words: sheep behavior detection, YOLO Master, Fourier angle alignment, efficient multiscale attention, image counting

Foundation item: Xi'an Municipal Science and Technology Bureau, Xi'an Municipal Science and Technology Plan Project (Agricultural Key Industry Chain Key Technology Tackling Project), Grant No. 24NYGG0023

First author: ZENG Yixiang, Email: zengyixiang@nwafu.edu.cn.

Corresponding author: WANG Meili, Email: wml@nwsuaf.edu.cn.

1 Introduction

Visual records of sheep provide information about feeding, stationary postures, and locomotion. A detector can locate visible animals and assign an activity label to each box, allowing the boxes to be summarized as total and class specific counts. Reviews of livestock motion monitoring, physiological sensing, and intelligent animal detection describe the role of visual information in livestock management [1,2,3]. The practical task studied here is to obtain behavior detections and visible animal counts from RGB images.

The agricultural literature increasingly addresses the difficulties that arise when animals overlap, occupy different image scales, or appear under variable illumination, and within this line of work DC YOLO explicitly studied occlusion, illumination variation, and class imbalance in housed Hu sheep [4], an improved YOLOv12 detector addressed occluded beef cattle [5], and a pig counting system evaluated instance segmentation together with counting errors [6]. These studies motivate the agricultural problem and the need to examine multiple operating measures.

Multiple animals in one image create two related evaluation requirements. Localization and class prediction determine detection accuracy, while the retained boxes determine the reported animal counts. Duplicate detections can inflate a count, and an incorrect class can change behavior counts even when the total remains correct. This study therefore evaluates detection and counting separately using public sheep videos recorded with mobile devices.

Behavioral monitoring also depends on the available observation modality. Edge based rumination recognition has been investigated using six axis signals and an improved MobileNet v3 model [7], and jaw movement sensing has been combined with feature selection for feeding related classification [8]. These measurements contain temporal or physical information that is absent from an isolated RGB image. A single image detector can learn visual correlates of running and walking labels, but movement speed and the duration of an activity require temporal observations. Consequently, the four labels examined here are treated as supplied image annotations, and image level behavior detection is distinguished from validated video activity recognition.

Single stage detection, exemplified by the original YOLO formulation [9], offers a direct route to predicting classes and boxes from an image. Agricultural adaptations often modify feature extraction, attention, or multiscale fusion. Lightweight cattle face recognition [10] and a recent cattle pose estimation study using attention and a higher resolution detection scale [11] illustrate related design directions, although neither task is equivalent to behavior detection. Beyond livestock, Flor YOLO combined frequency aware processing with architectural changes for carnation grading [12], and MHSF DETR combined heterogeneous feature extraction with context dependent fusion for cotton disease detection [13]. Their relevance is methodological, fine scale information and feature fusion are recurring concerns in agricultural perception. These studies motivate task specific evaluation of multiscale feature adaptation in sheep behavior detection.

YOLO Master introduces efficient sparse mixture of experts computation into a YOLO like detector [14]. Its adaptive backbone supports task specific investigation of multiscale feature fusion. Upsampled semantic features and higher resolution lateral features can encode different dominant structures. Fourier angle alignment addresses directional inconsistency in oriented remote sensing detection [15], and efficient multiscale attention recalibrates grouped features through cross spatial interactions [16]. Integrating these mechanisms at P3 offers a way to adapt feature fusion while retaining the existing resolution and detection scales.

This study tests a focused design hypothesis. At P3, estimating local direction before concatenation can reduce cross-level feature mismatch, while grouped context recalibration can refine the resulting high-resolution representation. The design changes the 80 × 80 pathway and retains the P4 and P5 prediction paths. Because the annotations describe individual images, detection average precision and image-level box counts are evaluated as separate outcomes.

This study makes three bounded contributions. First, it adapts the P3 feature-fusion pathway of YOLO-Master by applying reliability-weighted Fourier alignment to the upsampled semantic feature while retaining the lateral feature for concatenation. Second, it integrates EMA after the P3 fusion block to combine axis context and local spatial responses. The

joint configuration adds 3,746 parameters, corresponding to a 0.133% increase over the baseline. Third, it evaluates the baseline, single-module configurations, the joint configuration, and a projection control across three paired seeds, and reports detection AP and image-level count error as separate outcomes. The observed AP gains were positive for all three paired seeds, but their 95% confidence interval included zero. Accordingly, this study presents the modules as a task-specific, parameter-light adaptation with class-specific and counting-related effects, rather than as a statistically established accuracy improvement or a new state-of-the-art detector.

2 Dataset Construction and Preprocessing

2.1 Image Acquisition

All training, validation, and test images were derived from the public video dataset of Kelly et al. [18]. The original release contains 417 recordings from Pakistan and defines Grazing, Running, Sitting, Standing, and Walking as five activities. The source study reports acquisition using iPhone XS Max and Redmi Note 10 Pro mobile devices. The present study constructed a detection subset containing 2278 images and introduced its own four class vocabulary. Grazing was represented as Feeding, and Sitting and Standing were combined under Resting. This combination was made for the present annotations and was not part of the original five class definition.

Video screening and frame selection preceded bounding box annotation. The base extraction interval was five seconds. Videos showing faster movement underwent additional extraction rounds to retain more examples of changing postures. Consequently, the final subset includes both base samples and supplementary frames, rather than one uniformly spaced sequence from each recording. The retained image filenames and group assignments identify the exact subset used in the experiments. Training, validation, and testing use the same public source, and the test subset contains 240 images.

2.2 Data Annotation

Two annotators used LabelImg to draw bounding boxes and assign the study labels Feeding, Resting, Running, and Walking. A reviewer with expertise in animal science checked the annotations. Resting is an operational label for the combined Sitting and Standing categories, rather than a direct measurement of physiological rest. The retained subset contains 7386 valid boxes, comprising 4498 Feeding, 1446 Resting, 500 Running, and 942 Walking annotations. Figure 1 shows examples and the annotation interface, where the interface label run corresponds to Running.

Each annotation records a class index and the normalized box center, width, and height. Counts are derived by summing the valid boxes in each image, with separate sums for each class. This produces count targets directly from the detection annotations. The training images contain between 0 and 29 valid boxes per image, the validation images between 0 and 22, and the test images between 1 and 18. Image coordinates are converted to pixels for evaluation.

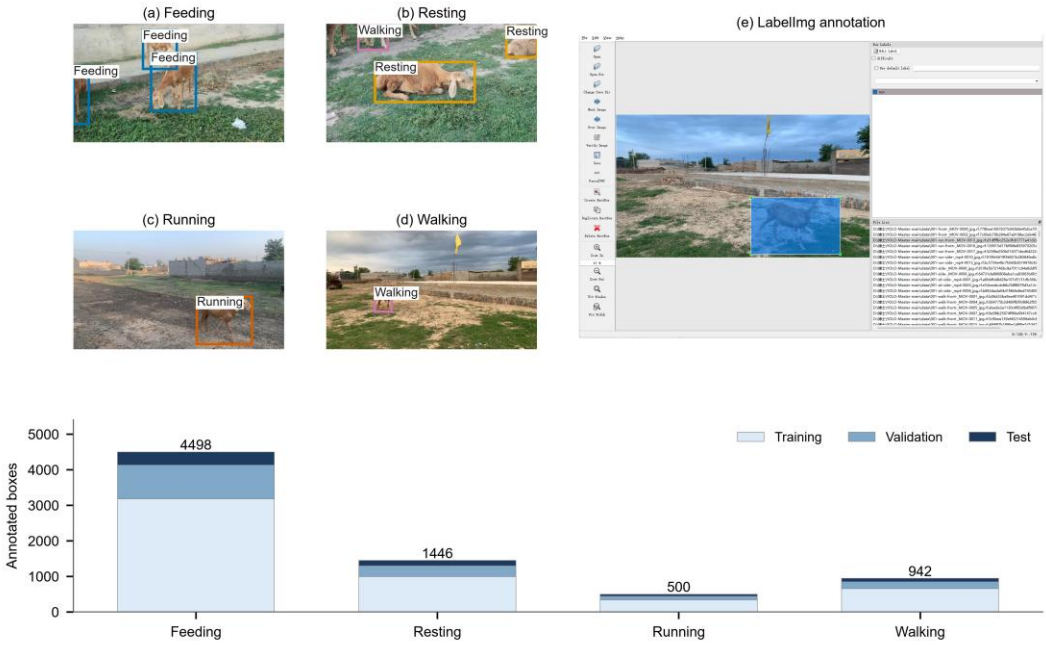


Fig. 1 Dataset examples, LabelImg annotation, and class distribution. Panels a to d show training examples containing Feeding, Resting, Running, and Walking, panel e shows the supplied LabelImg interface with a running sheep selected. Bars give valid annotation counts by subset. The interface tag run corresponds to Running in the experimental class vocabulary.

2.3 Dataset Partitioning and Augmentation

The retained images were organized into 239 filename groups. A group key was formed by removing the image extension and export hash, removing the numeric extracted frame suffix, and retaining the preceding filename prefix ending in mp4, mov, avi, or mkv. Matching was case insensitive. All images sharing a key were assigned together. The grouping script used random seed 42 and evaluated 12000 candidate assignments toward training, validation, and test proportions of 70, 20, and 10 percent. Selection balanced image counts and class box counts without using model predictions. The supplied image manifest and grouping script reproduce the retained membership and split.

The final split contains 165 training groups, 48 validation groups, and 26 test groups, with 1580, 458, and 240 images. No group key or exact image hash crosses these subsets. The 239 groups describe retained filename prefixes. A complete mapping from these prefixes to the 417 original video identifiers is unavailable in the retained files, so the group counts are not interpreted as a verified count of independently sourced original recordings. Table 1 reports the exact image and annotation counts. The evaluation measures separation under the recorded grouping rule within the public source.

Table 1 Dataset composition under the filename-derived grouped split

Subset	Groups	Images	Feeding	Resting	Running	Walking	Boxes
Training	165	1580	3178	992	336	658	5164
Validation	48	458	954	306	108	189	1557
Test	26	240	366	148	56	95	665
Total	239	2278	4498	1446	500	942	7386

Training used the same declared augmentation settings across formal configurations, mosaic probability 1.0, horizontal flip probability 0.5, translation fraction 0.1, scale range parameter 0.5, and HSV gains of 0.015, 0.7, and 0.4. Mosaic was disabled during the final ten epochs. Rotation, shear, perspective, vertical flip, mixup, and copy paste were disabled. Augmentations were applied within training after the split, validation and test images were not expanded with training augmentations. Images were letterboxed to the selected inference size and normalized according to the detector pipeline. Shared

augmentation settings are reported here, implementation specific preprocessing is retained in the cross family comparison.

2.4 Dataset Characteristics

Feeding accounted for 60.90 percent of valid boxes, and Running accounted for 6.77 percent. At the image level, Feeding occurred in 888 images, Resting in 422, Running in 500, and Walking in 882. An image can contain more than one class, so these image counts are not added to obtain the subset size. The test subset contained 366 Feeding, 148 Resting, 56 Running, and 95 Walking boxes. Classwise AP and macro averaged F1 complement aggregate AP when evaluating this class distribution.

Scale analysis used a reference input of 640 pixels. The image resize factor was the smaller of 640 divided by the original width and 640 divided by the original height. Reference box area was the original box width multiplied by its height and the square of this resize factor. The small, medium, and large bins used square root area thresholds of 32 and 96 pixels. The test subset contained 10, 348, and 307 boxes in these bins. The standard COCO definition based on original image area identified one small box. The geometric definition used for each result is stated explicitly. Animal age was not annotated.

3 Proposed YOLO Based Detection Method

3.1 Overall Network Architecture

Figure 2 shows the backbone, neck, and detection head. The design addresses two successive issues at P3. The upsampled feature U carries high-level semantics, while the lateral feature L preserves spatial detail. Their local activation patterns can differ around partial or overlapping sheep, and direct concatenation does not explicitly align them. FAAConcat therefore estimates reliable local direction and applies a residual correction to U before concatenation with the unchanged L. After the C3k2 transformation, grouped multiscale attention combines axis context with local spatial response. This order separates alignment from recalibration and confines both changes to the highest-resolution detection pathway. The rationale motivates the placement of the modules, while their measured effects are established by the controlled experiments.

The neck retains top-down and bottom-up fusion and modifies two consecutive P3 operations. FAAConcat replaces the original concatenation at layer 17. It receives U and L, each with dimensions of 128 × 80 × 80, and outputs Y with dimensions of 256 × 80 × 80. C3k2EMA replaces the original C3k2 block at layer 18, retains its feature transformation, and adds grouped attention with residual blending. Its output Z has dimensions of 64 × 80 × 80. The baseline P4 and P5 paths remain unchanged.

The head receives Z as P3 together with the unchanged P4 and P5 features, with dimensions of 128 × 40 × 40 and 128 × 20 × 20. The three levels use strides of 8, 16, and 32 and provide 6,400, 1,600, and 400 spatial prediction locations. Box decoding and non-maximum suppression convert predictions into retained boxes and behavior labels. The detector therefore gains no additional prediction scale, class, or regression loss.

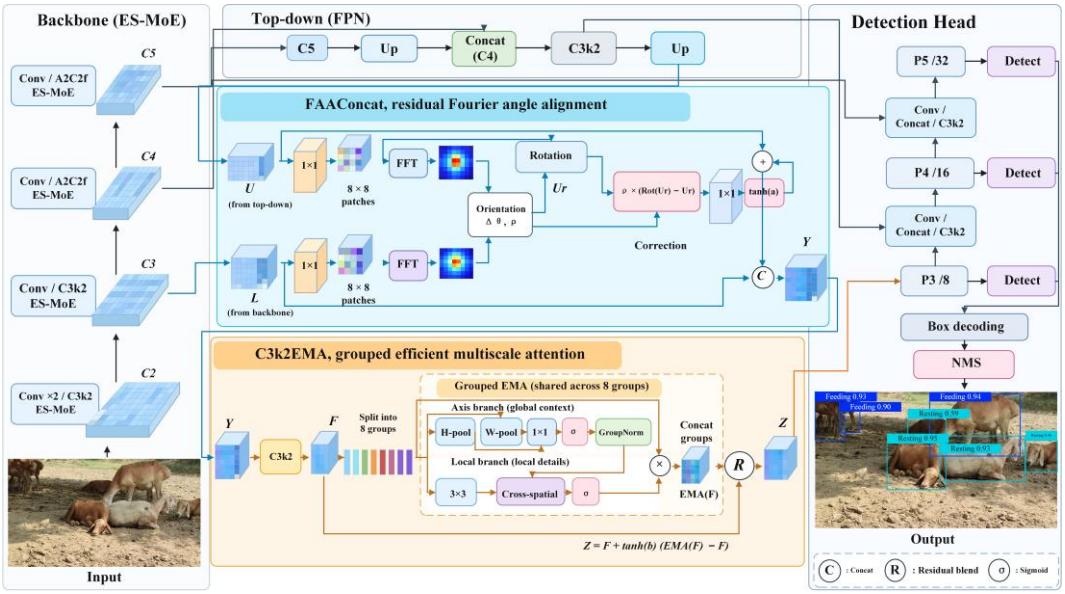


Fig. 2 Adapted YOLO Master architecture. Gray arrows denote the retained backbone and fusion routes, blue arrows denote the Fourier alignment pathway, and orange arrows denote the attention pathway. Arrow styles indicate data flow consistently. The cyan and orange panels expand FAAConcat and C3k2EMA. The eight colored slices indicate eight channel groups. U and L are the upsampled and lateral inputs, Y is their fused feature, F is the C3k2 output, and Z is the attention refined feature. C denotes concatenation, R denotes residual blending, and sigma denotes sigmoid activation. Input and output show the same training image with saved predictions.

3.2 Residual Fourier Angle Alignment at P3 Fusion

Fourier angle alignment was introduced for oriented remote-sensing detection [15]. Here, FAAConcat adapts its spectral direction cue to feature fusion for horizontal-box detection. It does not predict object angles or require angle labels. Instead, it uses local direction only to estimate whether the upsampled semantic and lateral features are geometrically consistent within each patch. The implementation uses non-overlapping patches, continuous doubled-angle spectral moments, reliability-weighted correction, and residual blending. Only the reduced semantic branch is transformed, and the original lateral feature remains available in the final concatenation.

At the selected P3 node, U and L each contain 128 channels at 80 × 80 resolution. Three bias-free 1 × 1 convolutions reduce both branches to d = 8 channels and project the final correction back to the 128-channel semantic feature. Each reduced feature is divided into non-overlapping 8 × 8 patches. Replicate padding is applied only when a spatial dimension is not divisible by eight. For each patch, the channel-averaged signal is transformed by a two-dimensional orthonormal FFT. Its squared magnitude is normalized to unit energy with a numerical lower bound.

$$q_1 = \sum_{u,v} p(u,v)\frac{f_u^2 - f_v^2}{\max(f_u^2 + f_v^2, \varepsilon)},$$
$$q_2 = \sum_{u,v} p(u,v)\frac{2 f_u f_v}{\max(f_u^2 + f_v^2, \varepsilon)} \quad (1)$$

Here, p denotes normalized spectral power at each frequency coordinate. The two frequency coordinates define the horizontal and vertical axes. The small constant prevents division by zero. The zero-frequency component is excluded from the directional basis. The doubled-angle moment represents an undirected axis, so its phase gives the local orientation and its magnitude gives the reliability of that estimate.

$$\theta = \tfrac{1}{2}\operatorname{atan2}(q_2, q_1 + 10^{-6}),$$
$$\rho = \min(1, \sqrt{q_1^2 + q_2^2 + 10^{-12}}) \quad (2)$$

Each patch produces an orientation and reliability for U and L. Their orientation difference is wrapped to the principal interval of an undirected axis by taking one half of atan2 applied to the sine and cosine of twice the difference. The affine matrix is built from this wrapped value and maps output sampling coordinates to input coordinates. This inverse-sampling convention fixes the rotation sign. Bilinear sampling uses border padding with corner alignment disabled. The rotated-minus-original semantic patch is weighted by the product of the two reliability values,

folded back to the feature grid, and projected to the original channel dimension.

$$U' = U + \tanh(s)W_o D,$$

$$Y = \mathrm{Concat}(U', L) \tag{3}$$

In Eq. 3, D is the folded correction and W_o is the restoring projection. The learnable scalar s is initialized to 0.01. The corrected semantic feature is concatenated with the unchanged lateral feature in the original channel order. When the spectrum is nearly isotropic, its low reliability suppresses the rotation-based correction. This reliability weighting makes the alignment conditional on directional evidence rather than forcing a correction on every patch. The formal experiments used FP32, including the FFT and grid-sampling operations.

FAAConcat adds 3,073 trainable parameters. Its projection cost scales with feature area, input channels, and the reduced channel count. After channel averaging, patchwise Fourier processing scales with feature area and the logarithm of the patch area. Rotation and folding also move feature data, so parameter count alone does not characterize runtime. Figure 2 shows the projections, spectral direction estimation, reliability weighting, residual correction, and concatenation with the unchanged lateral feature.

3.3 Residual Efficient Multiscale Attention

C3k2EMA first applies the retained C3k2 transformation to Y and produces a 64-channel feature F. Its attention stage adapts the grouped cross-spatial mechanism of efficient multiscale attention [16] to the P3 output. The 64 channels are divided into eight groups of eight channels. The same attention parameters are shared across groups, and the group dimension is folded into the batch dimension. This design exposes channel groups to the same context operator without assigning a separate parameter set to each group.

Within each group, horizontal and vertical average pooling encode axis-specific context. Their pooled descriptors are concatenated, passed through a shared 1 × 1 convolution, and converted into sigmoid gates. A parallel 3 × 3 convolution captures local responses. Global channel weights from each branch recalibrate the spatial response of the other branch. The two cross-calibrated responses are summed to form a sigmoid spatial attention map, which is applied to the original group feature. The operation preserves the channel count and spatial resolution.

The attended feature is combined with F through a learnable residual blend, shown as R in Figure 2. This preserves the original feature path while allowing the attention response to enter gradually. The scalar b is initialized to 0.1. Grouped attention adds 672 convolution and normalization parameters plus the residual scalar, for 673 parameters in total. Together, FAAConcat and C3k2EMA add 3,746 parameters, increasing the baseline from 2,813,236 to 2,816,982 parameters, or 0.133%.

$$Z = F + \tanh(a)(\mathrm{EMA}(F) - F) \tag{4}$$

In Eq. 4, a denotes the learnable blend coefficient, labeled b in Figure 2 and initialized to 0.1. The attention response is therefore introduced gradually while F remains the reference feature path.

3.4 Projection Control and Image Counting

The projection-only arm controls for trainable cross-level mixing. It retains FAAConcat's two reducing projections, restoring projection, residual scalar, input tensors, and output channel order, but replaces spectral direction estimation and patch rotation with the difference between the projected features. It therefore tests the alignment operation against a parameter-matched projection at the same fusion node. Because this control does not include EMA, it does not isolate the Fourier operation under the joint attention configuration.

$$Y_{\mathrm{proj}} = \mathrm{Concat}\,(U + \tanh(s)W_o(W_U U - W_L L), L) \tag{5}$$

All three projections remained trainable and active. The control matches FAAConcat in parameter count while replacing its spectral and geometric operations with projected feature differences. Together with the baseline, single module configurations, and joint configuration, it tests the contribution of trainable cross level feature mixing.

For each image, the counting interface sums retained detection boxes overall and by class. Ground truth counts are computed from the corresponding valid annotation boxes. The reported count errors describe visible annotated animals in individual images. They do not estimate unique identities over a

video, which requires temporal association such as that used by ByteTrack [19]. No tracker output enters the count metrics reported in this study.

4 Experimental Results and Discussion

4.1 Experimental Setup

All experiments ran on a remote GPU server whose computing environment is listed in Table 2. The experiment matrix contained five core configurations and five external detectors, each trained with seeds 42, 123, and 2026. Every run completed 100 epochs, early stopping was disabled with patience = 0. AdamW used an initial learning rate lr0 of 0.001 and a final learning rate fraction lrf of 0.01, corresponding to a final learning rate of 0.00001. Training included three warm up epochs, FP32 precision, deterministic execution, and an input size of 640 pixels. The core configurations used a physical batch size of 36.

Table 2 Computing environment and optimizer settings

Item	Setting
GPU	NVIDIA RTX 4090 D, 24 GB
Python / PyTorch	3.10.8 / 2.1.2+cu118
torchvision / NumPy	0.16.2+cu118 / 1.26.4
Operating system	Linux
Optimizer	AdamW
Initial learning rate (lr0)	0.001
lrf	0.01

Core runs started from a COCO pretrained YOLO Master checkpoint. For each seed, a shared four class initialization was saved and reused across core variants, retaining matched existing weights while initializing added modules. External detector families used their own compatible pretrained weights. The official YOLOv8n, YOLO11n, YOLO12n, and YOLO26n runs used an isolated Ultralytics 8.4.101 runtime [20]. YOLOv13 N used the supplied source implementation and native PyTorch scaled dot product attention when FlashAttention was unavailable. Runtime separation prevented unintended imports of the modified project package.

To use available GPU memory, physical batch sizes for YOLOv8n, YOLO11n, YOLO12n, YOLOv13 N, and YOLO26n were 92, 80, 64, 48, and 72, respectively. The nominal batch parameter nbs was 128. Integer gradient accumulation means that the resulting effective batches and optimizer step counts were not exactly identical across families. Cross family comparisons use the stated training budget with implementation specific batch sizes, initialization, and postprocessing.

All formal core configurations used an inference time expert threshold of zero to retain trained top k routing. The bias warm up learning rate was also zero.

The best checkpoint of each run was selected by its training implementation's validation criterion. A separate recorded comparison of mean common evaluator validation AP across the five core variants selected attention alone, baseline 59.07 percent, FAA 59.76 percent, EMA 60.79 percent, FAA+EMA 60.09 percent, and projection 60.67 percent. The joint configuration was a predefined experimental arm and is analyzed as such. Its test result did not overwrite the validation based architecture selection.

4.2 Evaluation Metrics

Bounding box average precision was calculated with pycocotools 2.0.10 using the COCO convention [21]. The primary metric averages precision over IoU thresholds from 0.50 to 0.95 in increments of 0.05 and over the four classes. AP50 uses an IoU threshold of 0.50. The common evaluator retained at most 100 detections per image. Predictions were exported using confidence 0.001, postprocessing IoU 0.7, and a maximum of 300 detections. Native postprocessing was retained for each model family, including the end to end path of YOLO26.

Fixed operating point classification statistics used confidence strictly greater than 0.25 and class correct greedy matches at IoU greater than 0.5. Precision and recall were calculated from true positives, false positives, and false negatives, and macro F1 averaged class specific F1 values,

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN},$$

$$F_{1,c} = \frac{2P_c R_c}{P_c + R_c} \tag{6}$$

The count for image i and class c was the number of retained predicted boxes assigned to c. Ground truth

counts were obtained from the corresponding valid annotations, including every test image. Total count mean absolute error and root mean square error were defined as follows,

$$\mathrm{MAE} = \frac{1}{N}\sum_i |\hat{n}_i - n_i|,$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i (\hat{n}_i - n_i)^2} \tag{7}$$

Behavior count macro MAE was the mean absolute class count error averaged equally over the four classes and all images. A total count can be correct even if several behavior labels are wrong, so the two errors were reported separately. Small object recall used the fixed 640 reference bins defined in Section 1.4 and the same fixed confidence and matching rule.

Means and sample standard deviations summarize three training seeds evaluated on the same test subset. Table entries place the sample standard deviation after each mean. Changes are calculated from unrounded values. Paired differences compare the joint configuration with the baseline at the same seed. A confidence interval for their mean is calculated using the Student t distribution with two degrees of freedom and the sample standard error of the three differences. It describes variation across training runs on this fixed test subset. Figure 3 presents precision recall curves using the aggregation specified in its caption.

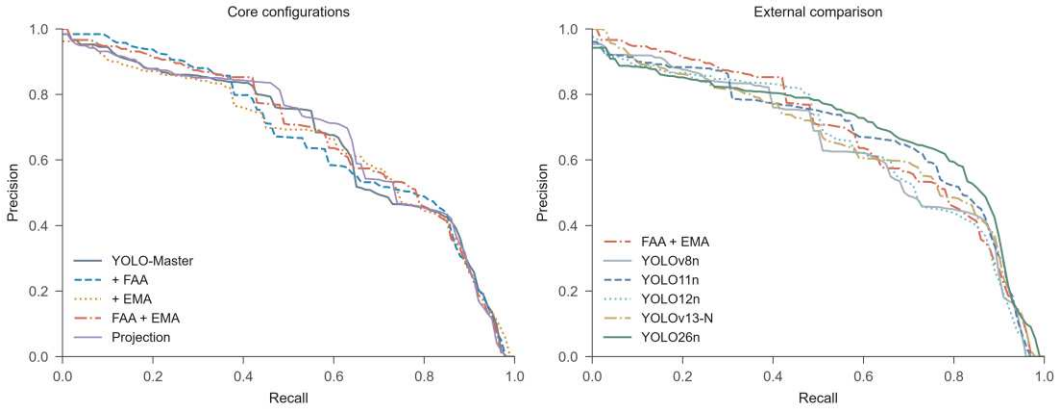


Fig. 3 COCO interpolated test precision to recall curves at IoU 0.50. Precision is averaged over four classes at each of 101 recall thresholds and then across three seeds, using all area predictions and maxDets=100. Left, core configurations. Right, external detectors.

4.3 Ablation Studies

Across the five core configurations, FAA+EMA achieved a mean test AP of 54.61% (SD 3.36), compared with 52.46% (SD 1.37) for YOLO-Master. The mean paired difference was 2.14 percentage points, with a sample SD of 3.00 percentage points and a 95% confidence interval from −5.30 to 9.59 percentage points. The interval includes zero, so the three-seed result does not establish a statistically significant mean AP increase.

Table 3 Controlled module ablation

Configuration	Parameters	Val AP (%)	Test AP (%)	Test AP50 (%)
YOLO-Master	2,813,236	59.07	52.46 (1.37)	66.30 (1.68)
+ FAA	2,816,309	59.76	52.89 (1.37)	66.14 (2.57)
+ EMA	2,813,909	60.79	52.27 (2.91)	65.10 (2.77)
+ FAA + EMA	2,816,982	60.09	54.61 (3.36)	67.13 (4.64)
+ Projection	2,816,309	60.67	54.45 (5.15)	67.15 (6.62)

The paired gains were +0.48, +5.61, and +0.35 percentage points for seeds 42, 123, and 2026, respectively. Thus, the positive mean was influenced substantially by one seed, although all three observed differences had the same direction. The joint model used 2,816,982 parameters, an increase of 3,746 parameters or 0.133% over the 2,813,236-parameter baseline.

The classwise results were uneven. Resting AP increased by 4.47 percentage points and Running AP

by 4.85 points, while Feeding increased by 0.43 points and Walking decreased by 1.18 points. These results locate the observed benefit mainly in Resting and Running, rather than showing uniform gains across all behaviors. The parameter-matched projection control reached 54.45% AP, only 0.16 percentage points below FAA+EMA, and used the same number of parameters as FAA alone. This small difference does not support attributing the full joint-model gain specifically to Fourier alignment. Taken together, the ablation results show a small parameter increase and positive paired AP differences within the tested YOLO-Master configurations, but the mean gain remains uncertain and the module-specific contribution is not fully separated by the current control.

EMA ranked highest on mean validation AP, whereas FAA+EMA ranked highest on mean core test AP. The joint configuration was a predefined experimental arm, validation based architecture selection and test evaluation are reported separately.

4.4 Comparison With Representative Lightweight Detectors

Under the reported evaluation protocol, FAA+EMA achieved a mean test AP of 54.61%, exceeding the measured AP of YOLOv8n, YOLO11n, YOLO12n, and YOLOv13-N. This comparison is specific to the tested checkpoints, dataset split, and evaluation settings.

Table 4 Comparison with representative lightweight detectors

Detector	Params (M)	Test AP (%)	Test AP50 (%)	Forward (ms)
YOLO-Master	2.813	52.46 (1.37)	66.30 (1.68)	24.72 (2.43)
+ FAA + EMA	2.817	54.61 (3.36)	67.13 (4.64)	28.23 (0.50)
YOLOv8n	3.012	51.08 (0.40)	64.00 (0.38)	7.56 (0.98)
YOLO11n	2.591	53.64 (4.28)	67.36 (4.46)	8.81 (0.29)
YOLO12n	2.569	50.37 (4.01)	64.41 (4.82)	15.47 (0.29)
YOLOv13-N	2.461	52.43 (3.09)	65.13 (2.87)	22.15 (0.17)
YOLO26n	2.505	55.94 (3.84)	69.17 (3.92)	13.56 (0.49)

YOLO26n achieved 55.94% AP with 2.505 million parameters and a forward latency of 13.56 ms. It therefore exceeded FAA+EMA in AP while also using fewer parameters and lower latency. The external comparison does not support an accuracy, model-size, or speed advantage for FAA+EMA over YOLO26n. Instead, it places the proposed configuration among the evaluated lightweight detectors and supports interpreting its contribution as a small-parameter adaptation within the YOLO-Master family.

All detector families were retrained on the same grouped dataset and scored by the common evaluator. Their pretrained weights, preprocessing, physical batch sizes, and native postprocessing followed the stated implementations. The comparison therefore characterizes these trained configurations under the recorded budget.

4.5 Classwise and Small Object Performance

Classwise AP identifies Resting and Running as the main contributors to the aggregate gain. Resting AP increased from 60.57 percent to 65.05 percent, a gain of 4.47 percentage points, and Running AP increased from 38.12 percent to 42.97 percent, a gain of 4.85 percentage points. Feeding AP increased from 71.39 percent to 71.82 percent, a gain of 0.43 percentage points, whereas Walking AP decreased from 39.77 percent to 38.59 percent, a reduction of 1.18 percentage points. Table 5 reports classwise means and changes calculated before rounding. Figure 4 compares the two configurations using grouped columns and across seed standard deviations.

Table 5 Classwise held-out detection performance

Behavior	Test boxes	Baseline AP (%)	Joint AP (%)	Δ AP (pp)
Feeding	366	71.39 (1.85)	71.82 (0.58)	+0.43
Resting	148	60.57 (3.34)	65.05 (0.94)	+4.47
Running	56	38.12 (5.38)	42.97 (11.27)	+4.85
Walking	95	39.77 (1.67)	38.59 (2.77)	-1.18

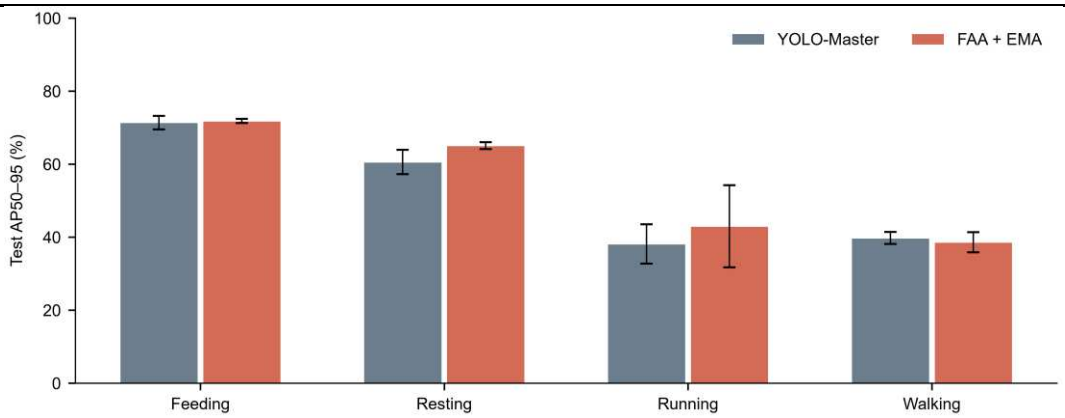


Fig. 4 Classwise test AP50 to 95 for YOLO Master and FAA+EMA. Grouped columns show the means across three matched training seeds, and error bars show the sample standard deviations of each configuration. The same seed pairs were used for both configurations.

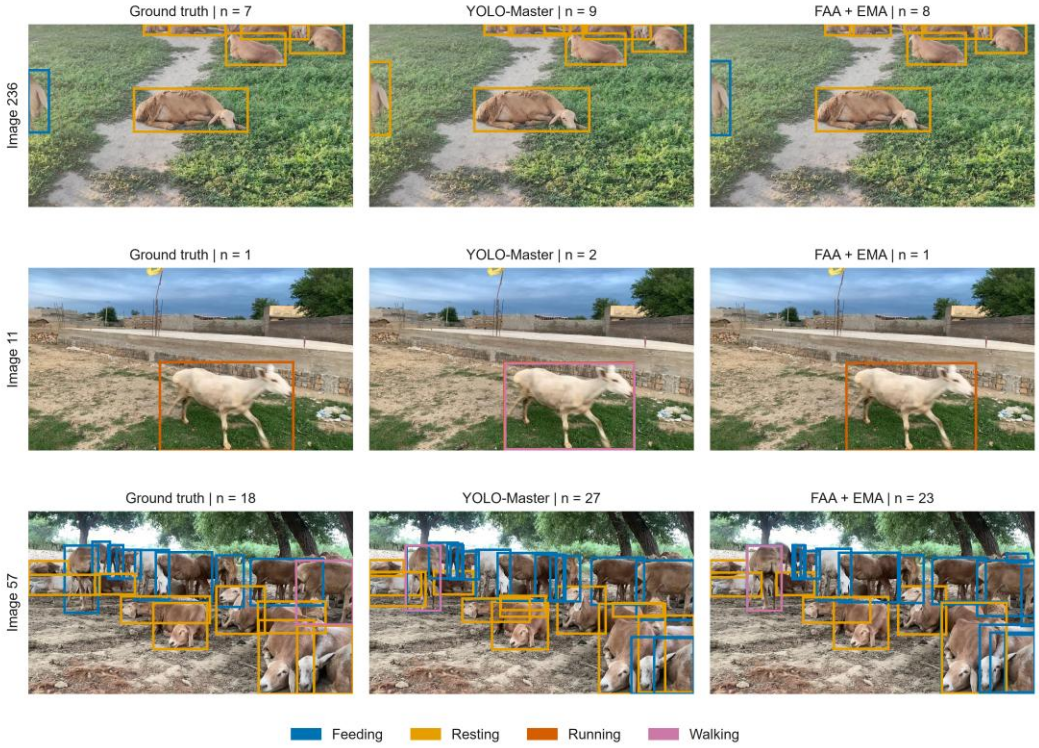


Fig. 5 Ground truth and seed 42 predictions on identical held out images from the public sheep activity dataset [18], at input size 640 and confidence greater than 0.25. Rows 1 to 2 are fixed annotation based examples, row 3 is a crowded scene diagnostic case. Colors identify behavior classes, and n denotes retained boxes.

Across ten reference defined small boxes, recall was 40.0 percent for both YOLO Master and the joint model. EMA achieved 50.0 percent, an increase of 10.0 percentage points. Under the standard COCO original pixel area definition, the test subset contains one small box, the two size definitions are explained separately.

Figure 5 shows crowded scenes, partial visibility, distant animals, and similar postures, which provide qualitative context for the aggregate test metrics.

Occlusion severity and illumination conditions were not annotated as quantitative comparison groups.

4.6 Grouped Holdout and Resolution Sensitivity

The 240 test images belong to filename groups withheld from training and validation within the same public collection. The split therefore evaluates transfer across the recorded groups of this source. It does not constitute testing on a separate farm dataset or independently acquired surveillance video. Resolution sensitivity was evaluated separately on validation images.

Increasing input resolution from 640 to 960 pixels changed baseline validation AP from 59.07 percent to 58.15 percent, a decrease of 0.91 percentage points, while small object recall increased from 60.69 percent to 68.24 percent, a gain of 7.55 percentage points. For EMA, AP decreased from 60.79 percent to 59.32 percent, a reduction of 1.46 percentage points, and small object recall increased from 64.78 percent to 74.84 percent, a gain of 10.06 percentage points. Figure 6 panel a compares AP and Figure 6 panel b compares recall using grouped columns. Both panels retain the three seed standard deviations, and recall uses the same 106 reference defined small validation boxes.

Increasing resolution raised small object recall while reducing aggregate AP for both examined configurations. Resolution sensitivity used the baseline and EMA configurations on validation images, formal test comparisons used 640 pixels.

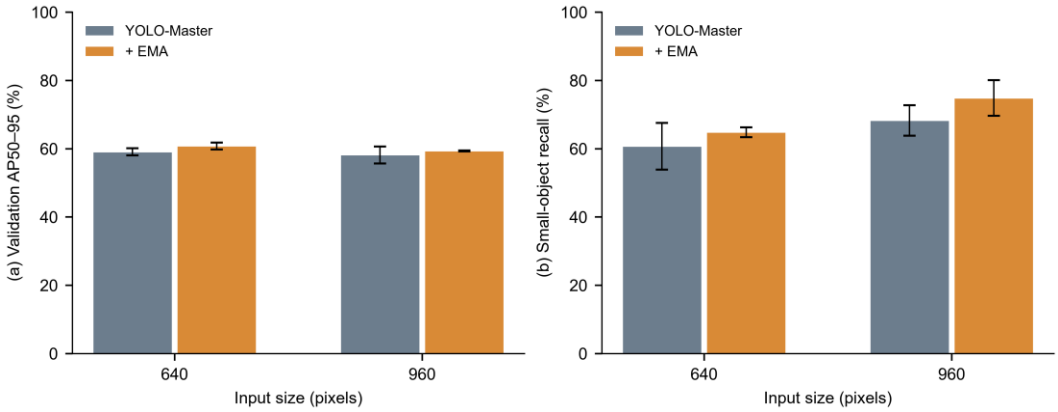


Fig. 6 Validation sensitivity to input resolution. Grouped columns compare YOLO Master and EMA at 640 and 960 pixels, a AP50 to 95 and b small object recall. Error bars show sample standard deviations across three seeds. Recall uses the same 106 small validation boxes defined at the 640 reference geometry.

4.7 Parameter Count and Network Forward Latency

All latency measurements used one NVIDIA RTX 4090 D with 24 GB memory. Each checkpoint was benchmarked in FP32 at batch size one using a fixed set of validation inputs, with 50 warm up iterations followed by 200 CUDA synchronized forward passes. The median latency was calculated for each seed, after which the mean and sample standard deviation of the three per seed medians were reported.

The joint model required 28.23 ms per forward pass, compared with 24.72 ms for YOLO Master, an increase of 3.51 ms, or 14.20 percent. FAA alone required 30.00 ms, EMA 27.26 ms, and the projection control 28.54 ms, increasing baseline latency by 5.28, 2.54, and 3.82 ms, respectively. The joint adaptation therefore combines a 2.14 percentage point mean AP gain with a 0.133 percent parameter increment and a 14.20 percent latency increase. Figure 7 displays AP and latency in separate aligned panels, parameter counts are reported in Tables 3 and 4.

YOLOv8n and YOLO11n required 7.56 and 8.81 ms per network forward pass, respectively, which reduced latency relative to YOLO Master by 17.16 and 15.91 ms. Together with the core model timings, these measurements separate parameter count from execution cost under the stated runtime.

The residual adaptation increases mean detection AP with 3,746 additional parameters. Its deployment cost is characterized here by network forward latency, complete camera pipeline timing is a separate measurement.

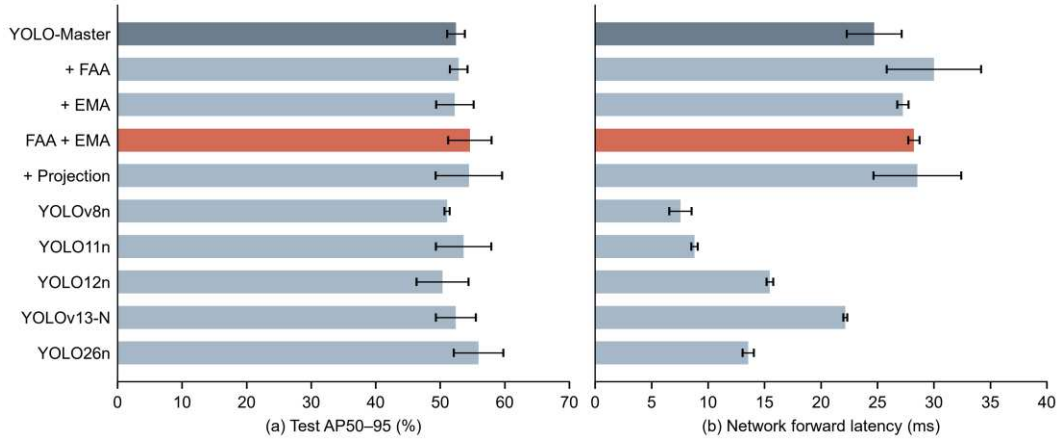


Fig. 7 Detection accuracy and network forward latency of ten configurations. Horizontal bars show a mean test AP50 to 95 and b mean batch one unfused FP32 network forward latency on RTX 4090 D. Error bars denote sample standard deviations across three seeds. Per seed latency is the median of 200 synchronized passes after 50 warm ups. Parameter counts are reported in Tables 3 and 4.

4.8 Counting Performance and Visualization

Detection AP and image-level counting measure different outputs. At confidence greater than 0.25, FAA alone reduced total count MAE from 0.549 to 0.401 animals per image, a 26.84% reduction. FAA plus EMA had the highest observed mean detection AP, while FAA alone had the lowest count MAE among the five core configurations. Reporting both measures is necessary because ranking detections across confidence and localization thresholds does not determine the number of boxes retained at one operating point.

Table 6 Fixed-threshold behavior and annotated-box count results

Configuration	Total MAE	Total RMSE	Behavior MAE	Macro-F1 (%)	Small recall (%)
YOLO-Master	0.549 (0.061)	1.131 (0.242)	0.278 (0.016)	69.09 (1.94)	40.0 (10.0)
+ FAA	0.401 (0.072)	0.819 (0.133)	0.248 (0.018)	69.68 (2.09)	40.0 (10.0)
+ EMA	0.576 (0.058)	1.283 (0.261)	0.295 (0.013)	67.65 (3.46)	50.0 (17.3)
+ FAA + EMA	0.574 (0.104)	1.239 (0.180)	0.306 (0.008)	67.78 (2.33)	40.0 (10.0)
+ Projection	0.436 (0.071)	1.011 (0.236)	0.259 (0.028)	70.99 (4.82)	43.3 (23.1)

Detection AP aggregates ranked predictions over confidence thresholds and localization criteria, whereas counting evaluates retained boxes at a fixed confidence threshold. Moderate confidence detections can affect ranking and retained counts differently. Behavior label substitutions can preserve the total count while changing per behavior errors. The observed metrics therefore support selecting configurations for the intended detection or counting objective.

Figure 5 compares identical held out public dataset images using seed 42 validation selected checkpoints. The same original pixels and all retained boxes are shown for ground truth, YOLO Master, and FAA+EMA. In Image 11, the baseline produced two detections for one annotated sheep, whereas FAA+EMA produced one. In Image 236, the baseline produced nine detections and FAA+EMA eight against seven annotations. Image 57 presents the crowded scene diagnostic case with 18 annotations and 27 versus 23 detections. These examples illustrate reductions in excess detections, aggregate accuracy is quantified over all 240 test images in Tables 3 to 6.

The evidence applies to the retained groups within the public mobile video collection. Testing on separately acquired farm recordings would assess transfer to another acquisition setting and would require corresponding activity and count annotations.

5 Conclusions

This study evaluated a parameter-light adaptation of the P3 pathway in YOLO-Master for four-class sheep behavior detection. The FAA+EMA configuration increased the parameter count from 2,813,236 to 2,816,982, a 0.133% increase, and achieved a mean test AP of 54.61% compared with 52.46% for the baseline.

All three paired seed differences were positive, but the mean gain of 2.14 percentage points had a 95% confidence interval from −5.30 to 9.59 percentage points and was influenced substantially by one seed. The mean AP gain therefore did not reach statistical significance. The observed classwise gains were concentrated in Resting and Running, while Walking decreased. FAA alone produced the lowest reported image-level count MAE among the core configurations, reducing it from 0.549 to 0.401, whereas FAA+EMA reached 0.574.

The projection control was only 0.16 percentage points below the joint configuration, so the present results do not isolate Fourier alignment as the cause of the joint AP outcome. YOLO26n also achieved higher AP with fewer parameters and lower latency. The evidence supports a parameter-light, directionally consistent, class-specific and counting-relevant adaptation within YOLO-Master. It also shows that the configuration with the highest observed detection AP need not minimize image-level count error, so model selection should reflect the intended output.

Declarations

Funding

This work was supported by the Xi'an Municipal Science and Technology Bureau through the Xi'an Municipal Science and Technology Plan Project (Agricultural Key Industry Chain Key Technology Tackling Project) (Grant No. 24NYGG0023).

Competing interests

Author HU Jian is employed by Xi'an Xiyangyang Biotechnology Co., Ltd., the lead organization of the funded project. The remaining authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Not applicable. No new animal experiments were performed in this study. We analysed only publicly available video recordings of sheep that had been released for research use by Kelly et al. (Data in Brief, 2024, 52: 110027).

Consent to participate

Not applicable. This study did not involve human participants.

Consent for publication

Not applicable. This study does not contain any individual person's data in any form.

Data availability

The dataset analysed during the current study is publicly available from Kelly et al. (Data in Brief, 2024, 52: 110027; https://doi.org/10.1016/j.dib.2024.110027). The bounding-box annotations and the filename-group split lists generated during the current study are available from the corresponding author on reasonable request.

Code availability

The training, evaluation, and counting scripts generated during the current study are available from the corresponding author on reasonable request.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by ZENG Yixiang and LIU Hongjie. The first draft of the manuscript was written by ZENG Yixiang,

and all authors commented on previous versions of the manuscript. Supervision, project administration and funding acquisition were carried out by WANG Meili and HU Jian. All authors read and approved the final manuscript.

References

[1] Wang, Z., Song, H., Wang, Y., et al.: Research progress and technology trend of intelligent Morning of dairy cow motion behavior (in Chinese). *Smart Agric. 4(2)*, 36–52 (2022). https://doi.org/10.12133/j.smartag.SA202203011

[2] Zhang, F., Zhou, M., Xiong, B., et al.: Research advances and prospect of intelligent monitoring systems for the physiological indicators of beef cattle (in Chinese). *Smart Agric. 6(4)*, 1–17 (2024). https://doi.org/10.12133/j.smartag.SA202312001

[3] Xiao, D., Lü, Y., Huang, Y., et al.: Research progress and future prospects of pig intelligent detection technology (in Chinese). *Smart Agric. 8(1)*, 86–103 (2026). https://doi.org/10.12133/j.smartag.SA202507048

[4] Li, X., Ji, R., Chang, H., et al.: DC-YOLO: a behavior detection model for Hu sheep addressing occlusion and illumination variations (in Chinese). *Smart Agric. 8(4)*, 127–147 (2026). https://doi.org/10.12133/j.smartag.SA202601021

[5] Liu, Y., Liu, L.: Beef cattle object detection method under occlusion environment based on improved YOLOv12 (in Chinese). *Smart Agric. 7(5)*, 182–192 (2025). https://doi.org/10.12133/j.smartag.SA202503018

[6] Zhang, Y., Zhou, S., Zhang, N., et al.: A regional farming pig counting system based on improved instance segmentation algorithm (in Chinese). *Smart Agric. 6(4)*, 53–63 (2024). https://doi.org/10.12133/j.smartag.SA202310001

[7] Zhang, Y., Li, X., Sun, Y., et al.: Real-time monitoring method for cow rumination behavior based on edge computing and improved MobileNet v3 (in Chinese). *Smart Agric. 6(4)*, 29–41 (2024). https://doi.org/10.12133/j.smartag.SA202405023

[8] Deng, X., Zhang, J., Liu, J., et al.: Dairy cow feeding-behavior recognition based on jaw-movement features optimized via model-performance feedback (in Chinese). *Smart Agric.* (2026). https://www.smartag.net.cn/EN/10.12133/j.smartag.SA202605016

[9] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91

[10] Han, Y., Qi, K., Zheng, J., et al.: Lightweight cattle facial recognition method based on improved YOLOv11 (in Chinese). *Smart Agric. 7(3)*, 173–184 (2025). https://doi.org/10.12133/j.smartag.SA202502010

[11] Liu, S., Li, Z.: Holstein cattle pose estimation via YOLOv11 with integrated SimAM and four-scale detection (in Chinese). *Smart Agric.* (2026). https://www.smartag.net.cn/EN/10.12133/j.smartag.SA202604052

[12] Li, C., Yang, J., Zhang, X.: Lightweight detection method for grading fresh cut Dianthus caryophyllus L. based on Flor-YOLO (in Chinese). *Smart Agric. 8(4)*, 85–99 (2026). https://doi.org/10.12133/j.smartag.SA202512007

[13] Shen, X., Zhang, Y., Jin, H., et al.: Multi-scale heterogeneous feature synergistic model for cotton leaf disease detection (in Chinese). *Smart Agric. 8(4)*, 204–216 (2026). https://doi.org/10.12133/j.smartag.SA202601027

[14] Lin, X., Peng, J., Gan, Z., et al.: YOLO-Master: MOE-accelerated with specialized transformers for enhanced real-time detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18440–18449 (2026). https://openaccess.thecvf.com/content/CVPR2026/html/Lin_YOLO-Master_MOE-Accelerated_with_Specialized_Transformers_for_Enhanced_Real-time_Detection_CVPR_2026_paper.html

[15] Gu, C., Chen, L., Gu, L., et al.: Fourier angle alignment for oriented object detection in remote sensing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42225–42235 (2026). https://openaccess.thecvf.com/content/CVPR2026/html/Gu_Fourier_Angle_Alignment_for_Oriented_Object_Detection_in_Remote_Sensing_CVPR_2026_paper.html

[16] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., et al.: Efficient multi-scale attention module with cross-spatial learning. In: *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10096516

[17] Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography 40(8)*, 913–929 (2017). https://doi.org/10.1111/ecog.02881

[18] Kelly, N.A., Khan, B.M., Ayub, M.Y., Hussain, A.J., Dajani, K., Hou, Y., et al.: Video dataset of sheep activity for animal behavioral analysis via deep learning. *Data Brief 52*, 110027 (2024). https://doi.org/10.1016/j.dib.2024.110027

[19] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., et al.: ByteTrack: multi-object tracking by associating every detection box. In: *Computer Vision – ECCV 2022*, pp. 1–21 (2022). https://doi.org/10.1007/978-3-031-20047-2_1

[20] Ultralytics: Models supported by Ultralytics. *Ultralytics Documentation* (2026). https://docs.ultralytics.com/models/

[21] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al.: Microsoft COCO: common objects in context. In: *Computer Vision – ECCV 2014*, pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48