

ORIGINAL RESEARCH

Open access

Automated Fish Maw Authenticity Detection Using a Lightweight YOLO Framework

Sofía Martínez^{1*}, Carlos Gómez¹, Lucía Navarro²

Abstract

Driven by the worldwide surge in fish maw commerce, authenticity verification has emerged as an urgent concern due to mounting adulteration incidents. Conventional approaches to verification—whether sensory-based, chemically oriented, or molecular—suffer from practical deployment constraints. The present study reframes fish maw genuineness assessment as a detection task within the object recognition paradigm and introduces RDT-YOLO, a streamlined detection model grounded in YOLO11n architecture. Addressing the twin difficulties of insufficient nuanced texture representation and processing overhead in this domain, we engineer hierarchically structured reparameterized modules for feature extraction that deploy reparameterization tactics to amplify multi-resolution texture discrimination. To counteract degradation of information arising during cross-scale feature merging, a Dynamic Adaptive Multi-Scale Pyramid Processing (DAMSPP) component is devised, integrating adaptive convolution dynamics that facilitate responsive feature synthesis. Concurrently, an Adaptive Task-Aligned Detection Head (ATADH) is formulated, fusing interactive task mechanisms with weight-shared convolutions to achieve parameter economy alongside precision gains. Moreover, Wise-ShapeloU, a novel loss formulation, is constructed by injecting Wise-IoU's focusing coefficient into the Shape-IoU framework, thereby sharpening detection fidelity through refined contour alignment. Benchmark experiments confirm that RDT-YOLO delivers precision of 91.9%, recall of 89.6%, and mAP@0.5 of 94% while slashing parameter count, storage footprint, and FLOPs by 75.6%, 73.8%, and 63.8%, respectively, against YOLO11s. Comparative trials versus YOLOv10s and YOLOv12s reveal mAP@0.5 elevations of 0.8 and 0.5 percentage points. This investigation furnishes an automated pipeline for verifying fish maw genuineness, bearing implications extending across food integrity applications.

Keywords Food safety, Deep learning, Fish maw authenticity detection, YOLO11

*Correspondence:

Sofía Martínez

sofia.martinez@upm.es

¹ Department of Computer Vision for Seafood Authentication, Faculty of Agricultural Sciences, Polytechnic University of Madrid, Madrid, Spain

² Department of Lightweight Deep Learning for Food Quality, Faculty of Agriculture, University of Barcelona, Barcelona, Spain

Introduction

According to documentation from the FAO, global aquaculture output has registered substantial expansion [1], positioning China at the center of consumption for premium desiccated marine commodities—shark fin, abalone, and fish maw among them [2]. Fish maw denotes a desiccated marine derivative manufactured through artisanal curing of swim bladders sourced from particular fish varieties. The product boasts abundant proteinaceous content along with vitamins and micronutrients, conferring therapeutic attributes that encompass styptic and inflammation-moderating effects [3, 4]. Concurrent with escalating market appetite, manufacturers have turned to a broadening spectrum of source species [5]. Valuation in the fish maw trade hinges overwhelmingly on species identity and qualitative grading; discrepancies in pricing

across species boundaries and quality strata have spawned pervasive practices of adulteration, species substitution, and deceptive labeling [6–8]. Across the globe, alimentary fraud has intensified into a matter of grave concern [9]. These circumstances underscore the pressing imperative for robust fish maw verification methodologies.

Legacy authentication protocols for fish maw products rest chiefly upon three pillars: organoleptic examination, instrumental chemical profiling, and biomolecular assays. Organoleptic judgment, while demanding seasoned proficiency, remains intrinsically prone to individual bias and nontrivial misclassification rates [10]. Instrumental chemical profiling spans spectroscopic modalities [11] alongside chromatographic separation coupled with mass spectrometric detection [12]; biomolecular strategies leverage DNA-centric workflows encompassing PCR amplification [13], barcode-based species assignment, and next-generation sequencing [14]. Notwithstanding the analytical fidelity these instrumental and molecular pipelines afford, their translation into routine practice is hampered by protracted turnaround, prohibitive expenditure, dependency on dedicated hardware, and the necessity of highly trained operators [15, 16]. A paramount objective thus lies in engineering imaging-derived, nondestructive, field-deployable verification technologies suited to fish maw screening.

Breakthroughs propelled by contemporary artificial intelligence have reshaped food-science applications of deep neural networks [17]. In the visual recognition arena specifically, deep architectures have upended paradigm object detection frameworks, cementing neural-network-based formulations as the prevailing methodology. Current detection frameworks bifurcate into two-stage schemes—R-CNN [18], Fast R-CNN [19], and Faster R-CNN [20]—and single-stage alternatives encompassing SSD [21], RetinaNet [22], together with successive generations of the You Only Look Once lineage [23–27]. Two-stage configurations pursue elevated accuracy via serialized region nomination and subsequent classification, whereas single-stage formulations prioritize inferential throughput and suitability for time-sensitive contexts.

Within alimentary detection settings, YOLO enjoys pervasive adoption owing to its trifecta of swift inferential passes, lean parametric footprints, compact model realizations, and dependable precision. The model family's adaptability finds validation across a breadth of food-authentication undertakings. Illustratively, Geng *et al.* [28] advanced a micrographic deep-learning methodology for detecting fishmeal adulterants, enlisting Mobilenetv2 as the qualitative discriminator and coupling it with YOLOv3-Mobilenetv2 to accomplish ingredient-level identification with notable exactitude. Kong *et al.* [29] documented an AM-YOLO-driven authentication framework for *Astragalus membranaceus*, elevating both recognition fidelity and inferential latency in distinguishing authentic from spurious botanical specimens. Jubayer *et al.* [30] enlisted YOLOv7 to perform apicultural pollen enumeration and taxonomic sorting, thereby strengthening honey provenance verification. Zhang *et al.* [31] engineered a pruned Faster-YOLO variant anchored in YOLOv8n, embedding grouped-convolutional hybrid-attentional elements to tackle rice-variety adulteration challenges. Jose *et al.* [32] formulated a YOLOv8-based spice-classification pipeline incorporating variably-kerneled convolutional architectures, materially boosting discriminative precision among visually confusable spices. Collectively, these inquiries substantiate both the potency and the broad applicability of YOLO-centric strategies across the food authentication landscape.

Nonetheless, the challenge of discriminating fish maw species embodies idiosyncratic hurdles that remain insufficiently confronted despite YOLO's attested success in diverse alimentary classification contexts. Departing from foodstuffs bearing overt visual signatures, specimens of fish maw drawn from divergent species present markedly congruent gross morphologies, diverging solely in delicate surface textural nuances and patterning subtleties—a reality demanding extraction capabilities of heightened sophistication. Compounding this, the pronounced dimensional and configurational heterogeneity engendered by varied curing and processing regimes compels the adoption of robust multi-granular detection apparatuses. Practical deployment exigencies—such as field-level inspections conducted within wholesale distribution hubs or retail premises—further require models sufficiently lightweight to execute upon portable computational substrates while forfeiting minimal detection acuity. These interlocking requisites crystallize the imperative for bespoke detection frameworks expressly architected to meet the particular demands of fish maw genuineness screening.

In direct response to the foregoing challenges, the present contribution puts forward RDT-YOLO, an augmented YOLO11-anchored model fine-tuned for the distinctive demands of fish maw verification. Diverging from off-the-shelf

detection models, RDT-YOLO instantiates a coherent, systematically designed paradigm calibrated to the singular attributes of the fish maw authentication task—namely, near-imperceptible textural divergences, pronounced scale variability, and the necessity of lightweight operational profiles. The overarching methodological scaffold is depicted schematically in **Figure 1**.

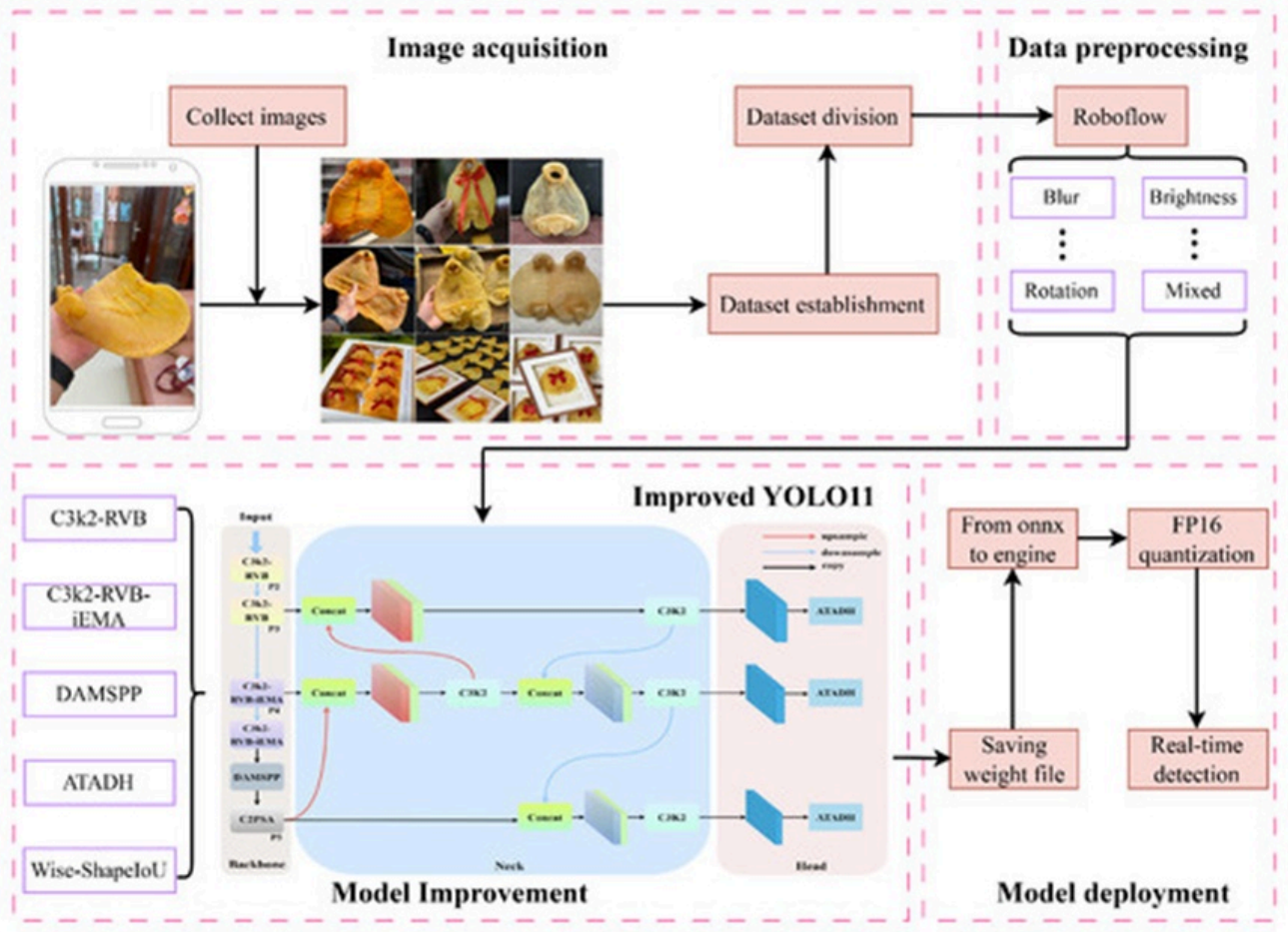


Figure 1. Fish maw authenticity detection framework.

The salient contributions and novel elements of this investigation may be summarized as follows:

- To the best of our knowledge, this study marks the inaugural deployment of deep-learning-grounded object detection technology to the fish maw authenticity verification problem, effecting a transformation from a classical analytical-chemistry pipeline into a visual-recognition undertaking and furnishing an extensive image repository concentrating on taxonomically confusable species pairs.
- We architect a stratified dual-phase feature encoding hierarchy, comprising C3k2-RVB modules within superficial strata and C3k2-RVB-iEMA counterparts within deeper strata. Through the substitution of standard SE attention with our purpose-built iEMA module inside a RepViT block, spatial-contextual awareness is amplified via cross-spatial representational learning expressly optimized for the fine-texture discrimination demands inherent to fish maw examination.
- We conceive the DAMSP module, which integrates a pioneering Adaptive Inception Convolution (AIConv) operator alongside a Dynamic Kernel Weighting (DKW) gating paradigm. This construct adaptively apprehends fish maw surface-texture signatures across a continuum of granularities through dynamic feature amplification, thereby

transcending the representational rigidity imposed by static pooling regimes characteristic of traditional SPPF implementations.

- We develop an Adaptive Task-Aligned Detection Head (ATADH) that orchestrates a synergistic interplay among task-decomposition principles, weight-shared convolutional layers, and DyDCNv2 mechanisms, yielding productive harmonization between the classification and spatial-regression sub-tasks while upholding a notably economical parametric profile.
- We fuse the Wise-IoU V3 modulation coefficient into the Shape-IoU formulation, yielding a composite Wise-ShapeloU regression objective explicitly engineered to accommodate the heterogeneous sample-quality distribution encountered in authentic fish maw screening environments, thereby attaining superior bounding-box convergence characteristics.

Materials and Methods

Data acquisition

The image corpus underpinning this investigation was predominantly assembled from a pair of sites: the Shanghao Jiao retail establishment in Shantou City, Guangdong Province, China, together with Jiexun Aquaculture Co., Ltd., headquartered in Raoping County, China. A multi-stream collection methodology was pursued to strengthen the heterogeneity and representational breadth of the assembled samples. The finalized fish maw image repository spans five principal groupings: Red-mouth fish maw, White croaker fish maw, Annan fish maw, Douhu fish maw, and Duck-tongue shaped maw.

Marked divergence characterizes the biological provenance of these distinct fish maw varieties. Red-mouth fish maw traces its lineage to a notably convoluted array of progenitor species, chiefly encompassing taxa such as *Protonibea diacanthus* and *Megalonibea fusca*. The inherent sparseness and exceptional uncommonness of these piscine sources in the wild confer upon Red-mouth fish maw a lofty commercial valuation, positioning it among the most esteemed grades of fish maw. By contrast, the donor organisms behind the remaining four categories enjoy comparatively widespread availability: Annan fish maw derives from *Otolithoides biauritus*, falling within the *Otolithoides* genus; White croaker fish maw and Duck-tongue shaped maw alike represent marine-sourced maws originating from members of the *Pseudolithus* clade; whereas Douhu fish maw stems from *Galeoides* stock. Attributable to the relative plenitude of their biological sources, these latter maw types occupy a markedly lower market-price tier vis-à-vis Red-mouth fish maw. Yet their outward physical presentation proves exceedingly congruent, rendering them frequent instruments in the fraudulent imitation of Red-mouth fish maw specimens.

Precisely because inter-category morphological resemblances are so pronounced and marketplace adulteration so subtle, the assembly of a repository capable of faithfully resolving these minute divergences assumes foundational significance. In pursuit of this objective, image capture was conducted via a Canon EOS 6D body (Canon Inc., Tokyo, Japan), yielding frames of 4032×3024 pixel granularity, thus guaranteeing high-resolution documentation. To safeguard variance and ecological validity across the dataset, photographic recording spanned manifold lighting regimes, specimen orientations, qualitative strata, and acquisition perspectives, embracing scenes with solitary and clustered objects as well as visually congested background settings. **Figure 2** exhibits archetypal specimen images that make manifest the signature hallmarks and subtle inter-class distinctions separating the five fish maw designations.

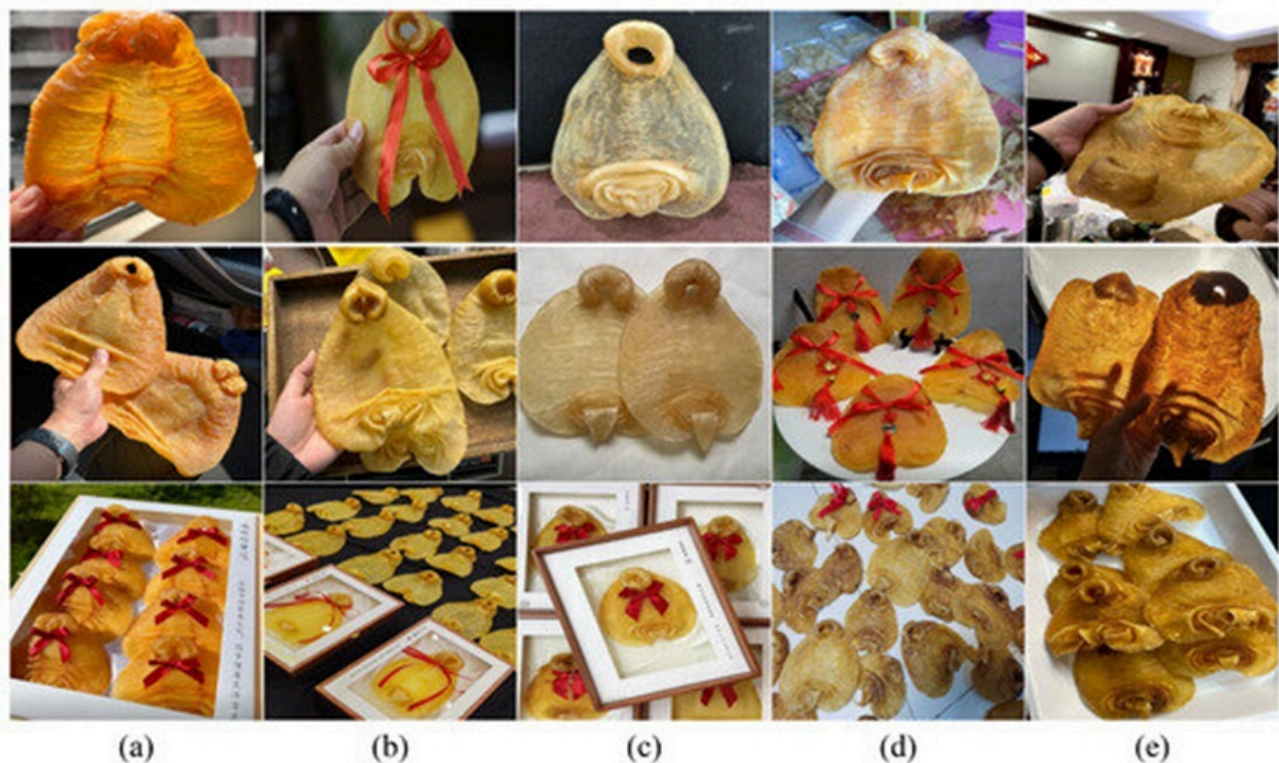


Figure 2. Samples from a portion of the fish maw dataset: (a) Red-mouth fish maw (ZJCZ), (b) Annan fish maw (AN), (c) White croaker fish maw (BM), (d) Douhu fish maw (DH), and (e) Duck-tongue shaped maw (YL).

Data preprocessing

A regimented preprocessing sequence was instituted to preserve the overall caliber of the dataset. An initial quality-auditing pass was carried out through the gathered fish maw imagery, culling any instances compromised by motion-induced smearing, aberrant exposure values, or insufficient pixel resolution. Subsequently, every retained frame was subjected to dimensional standardization, undergoing resizing to a uniform resolution of 640×640 pixels. During the annotation stage, meticulous bounding-box markup was undertaken through the Roboflow environment for each of the five target categories. In a concluding step, the annotated body of data was stochastically allocated following a 7:2:1 ratio, thereby generating training, validation, and testing partitions.

Data augmentation

A preliminary training regimen of 200 epochs was administered to the YOLO11n architecture to gauge baseline capability. The outcomes revealed that the ZJCZ class attained superior recognition fidelity, whereas the BM and DH groupings exhibited comparatively diminished scores. Augmentation tactics were marshalled to counteract this performance asymmetry [33].

The suite of transformations embraced stochastic rotation bounded within $\pm 15^\circ$, luminance perturbation spanning a $\pm 25\%$ envelope, and Gaussian blur whose kernel radius extended as far as 1.5 pixels. Collectively, these interventions approximate the influence of shifting observational angles, fluctuating ambient illumination, and variable optical sharpness encountered in practice. Each frame drawn from the original training corpus generated a single augmented facsimile, enlarging the training contingent from 1749 to 3498 instances, whilst the validation and testing partitions remained unaltered. **Figure 3** illustrates representative augmentation outputs. **Table 1** enumerates specimen tallies pre- and post-augmentation.

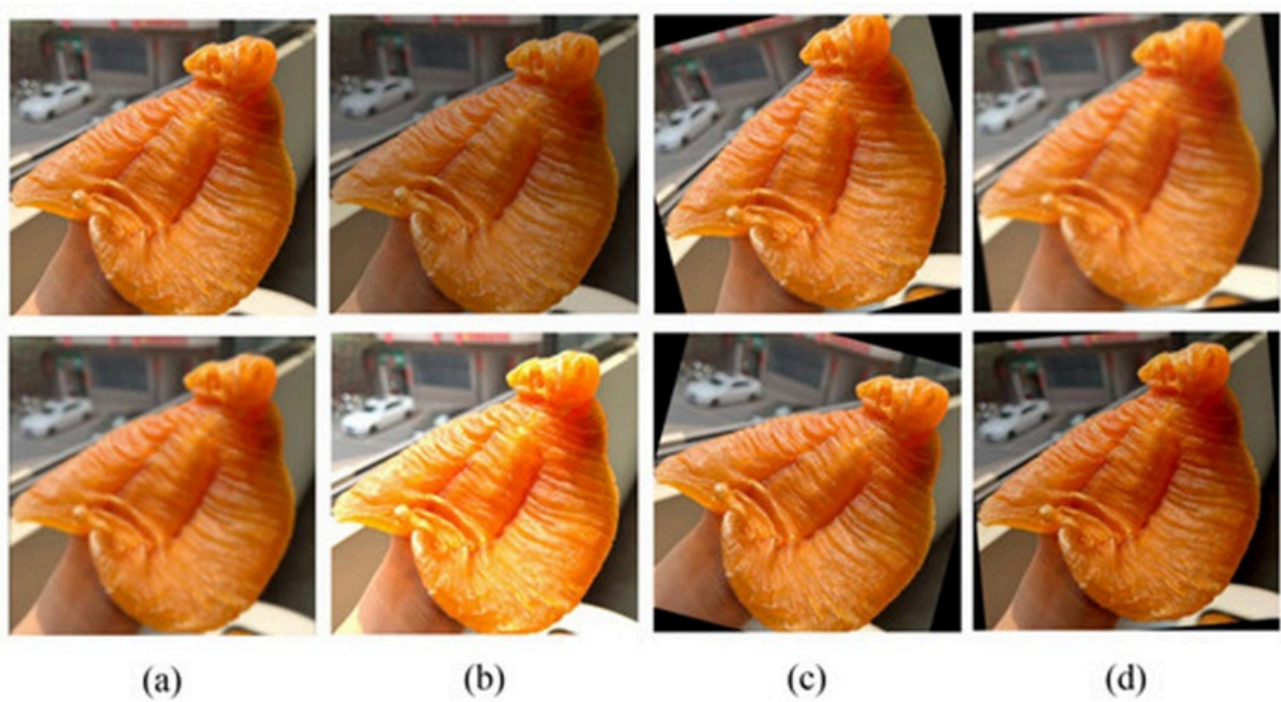


Figure 3. Illustration of image enhancement strategies: (a) original and blur up to 1.5px, (b) brightness adjustment between -25% and $+25\%$, (c) rotation between -15° and $+15^\circ$, and (d) mixed augmentation.

Table 1. Distribution of the fish maw dataset before and after data augmentation.

Categories	Before data augmentation	After data augmentation			
		Training set	Validation set	Testing set	Total
AN	382	544	74	36	654
BM	346	490	71	30	591
DH	363	506	71	39	616
YL	453	618	96	48	762
ZJCZ	950	1340	184	96	1620

RDT-YOLO network structure

RDT-YOLO is erected atop the YOLO11 scaffold and purpose-tailored to surmount difficulties tied to impoverished fine-texture representation mining, bottlenecked cross-resolution feature amalgamation, and superfluous parametric overhead in the sphere of fish maw genuineness screening. The C3k2-RVB and C3k2-RVB-iEMA building blocks have been contrived to displace the native backbone C3k2 constituent of YOLO11, achieving tangible compression of model parameter count while markedly sharpening sensitivity to subtle surface textural signatures characteristic of fish maw—thereby rectifying recognition failures originating from textural resemblance and cluttered scenery. To rise above what static SPPF pooling regimes can offer, the DAMSPP unit has, in parallel, been crafted. This construct adaptively registers fish maw surface cues across a continuum of granularities by way of dynamic feature accentuation and hierarchical representation extraction, substantially elevating feature descriptive power.

Multi-resolution representational abstractions extracted from the backbone cascade forth into the neck subnetwork, where cross-scale feature synthesis is accomplished. The resulting triplet of feature planes is subsequently routed toward the detection terminal. The ATADH has been formulated to supersede the stock detection head, achieving pronounced parameter shrinkage via task-interactive gating while efficiently disambiguating feature-encoded intelligence to localize targets spatially, assign their semantic labels, and quantify detection certitude. Concomitantly, the Wise-ShapeloU cost function has been instantiated to sharpen localization fidelity. **Figure 4** delineates the structural schema of the optimized RDT-YOLO, with its nucleus constituents unpacked in the sequel.

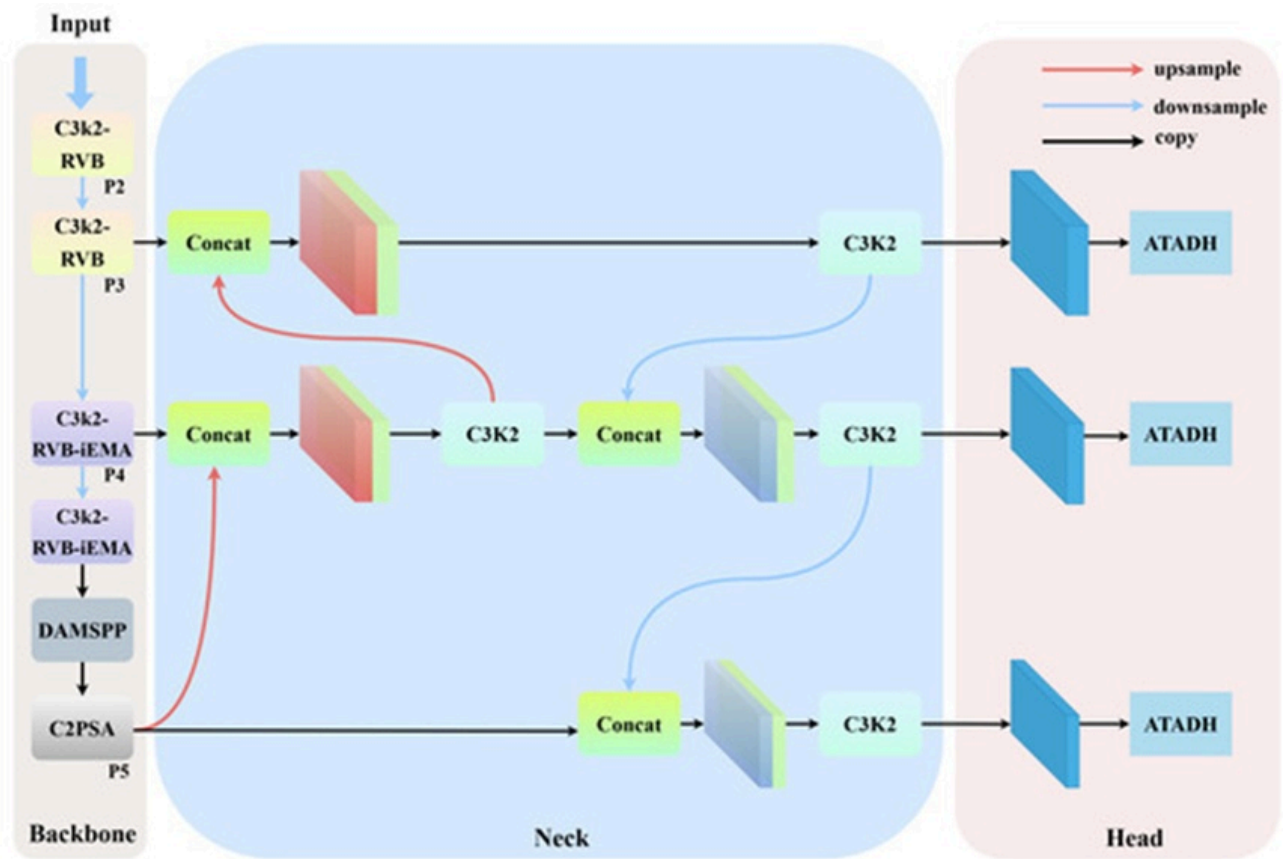


Figure 4. Model structure diagram of RDT-YOLO.

Reparameterized feature extraction module

Discriminating genuine from counterfeit fish maw rests, at its core, on disentangling species whose outward morphology appears near-identical by capitalizing on barely perceptible dissimilarities in surface grain. As visualized in **Figure 5a**, legacy C3k2 modules are hamstrung by two decisive deficiencies: first, their bottleneck topology sacrifices high-frequency detail through channel compression, thereby blurring the boundaries among confusable textures; second, the omission of spatially adaptive attention precludes selective accentuation of diagnostic textural signatures. These flaws translate squarely into erroneous classification verdicts whenever the model is presented with visually proximate fish maw classes.

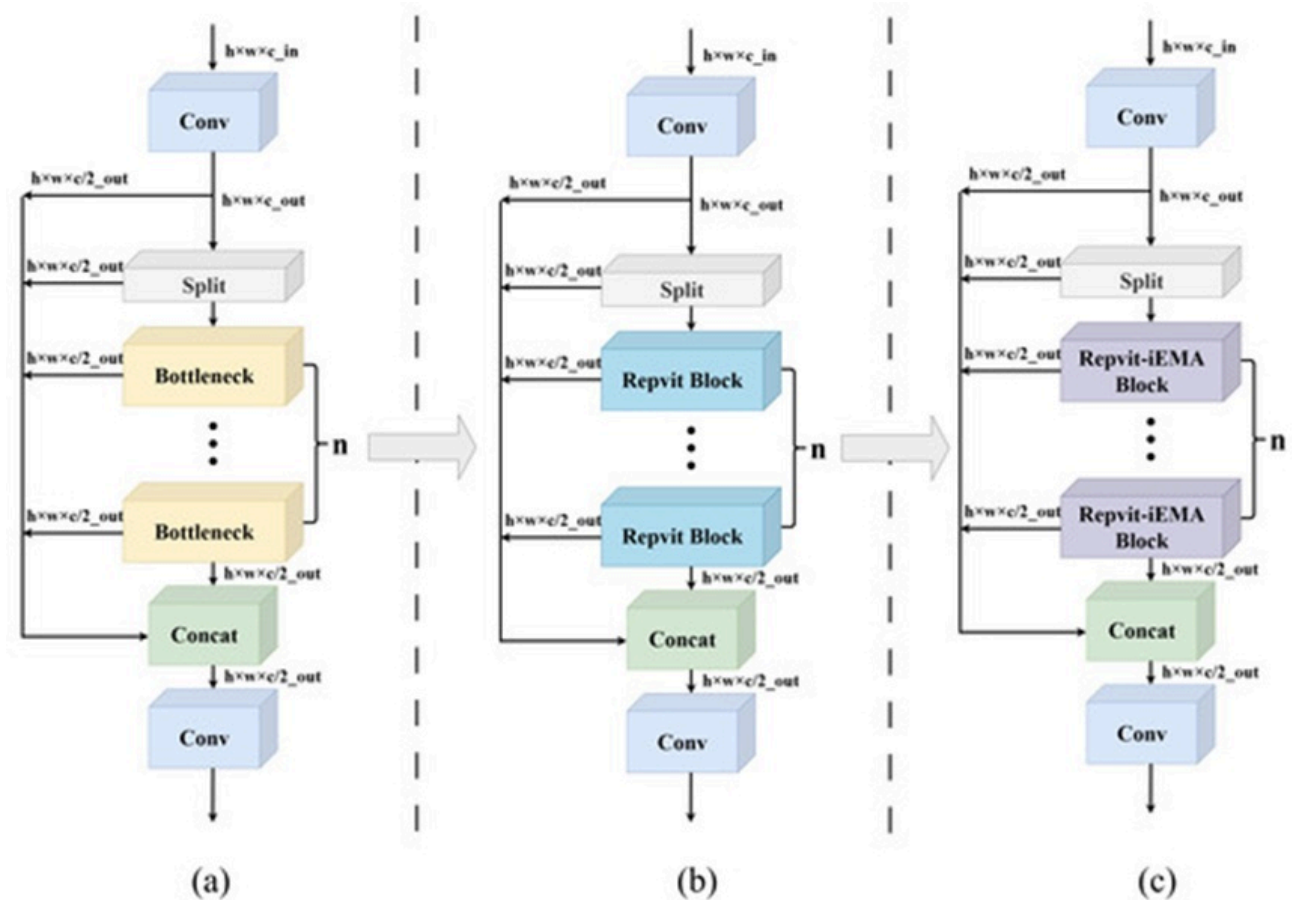


Figure 5. Structure of the C3k2-RVB-iEMA module: (a) C3k2 module, (b) C3k2-RVB module, and (c) C3k2-RVB-iEMA module.

In answer to these impediments, we advance a hierarchically bifurcated architecture that materializes incremental texture assimilation. Within superficial strata, the RepViT blueprint [34] gives form to the C3k2-RVB module, illustrated in **Figure 5b**, which yields reparameterization to conserve fine-grained fidelity without sacrificing execution economy. Within deeper tiers, we fashion the C3k2-RVB-iEMA module, rendered in **Figure 5c**, by embedding our custom-engineered iEMA unit, which seizes orientation-specific textural motifs through cross-spatial relational learning enacted via paired 1D pooling operators aligned with the width and height axes.

iEMA feature enhancement module

When deployed for fish maw genuineness screening, off-the-shelf attention formulations grapple with the intricacies posed by elaborate surface textural patterns. To transcend this shortfall, we present the integrated Efficient Multi-scale Attention (iEMA) feature augmentation module, depicted in **Figure 6**, which marries EMA-style attention with an inverted residual design philosophy.

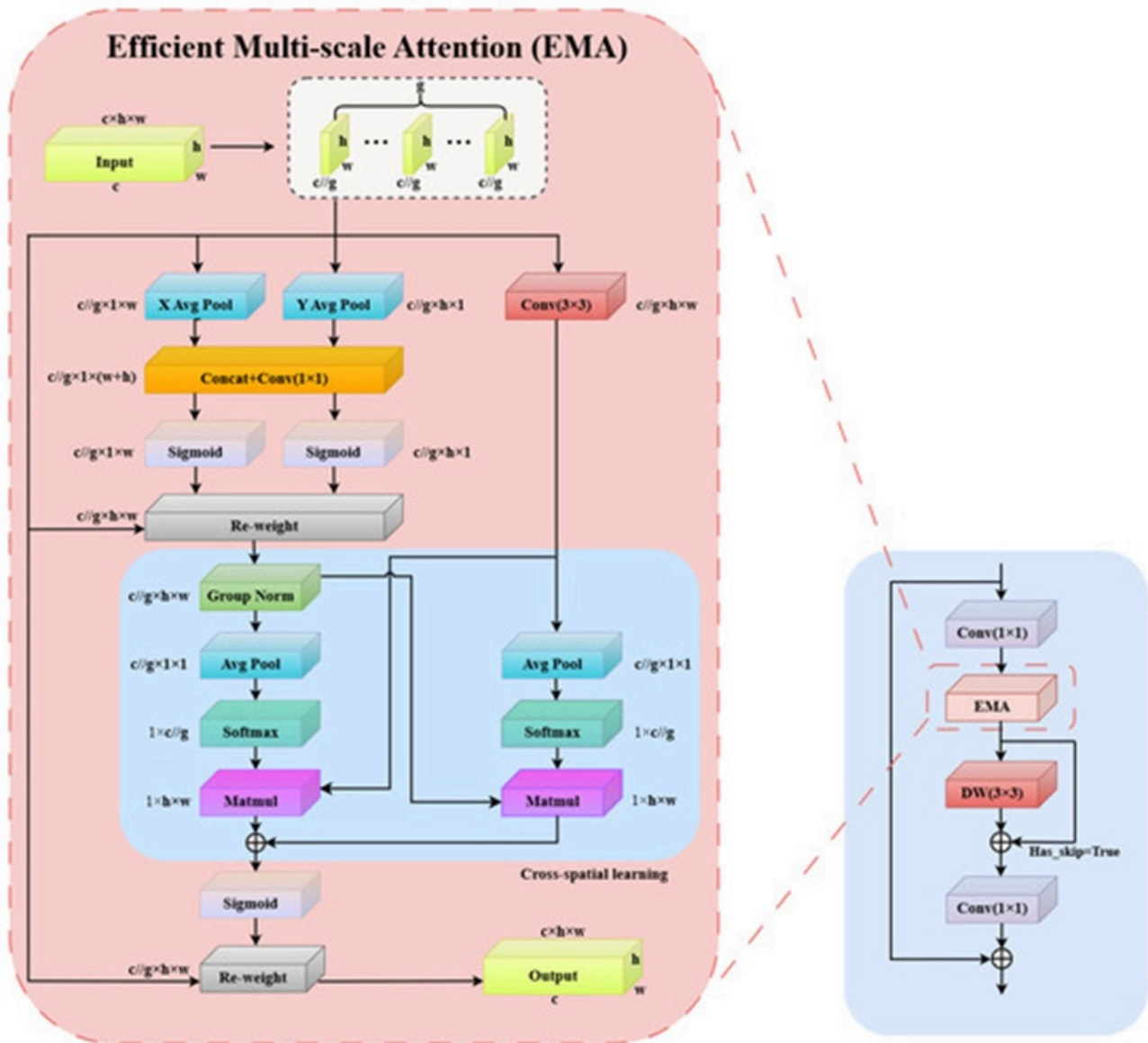


Figure 6. Structure of the iEMA module.

The EMA [35] operator encodes spatial interdependence by applying 1D global average pooling along both the horizontal and vertical dimensions in parallel:

$$z_c^H = \frac{1}{H} \sum_{h=0}^{H-1} x_c(h, w) \quad (1)$$

$$z_c^W = \frac{1}{W} \sum_{w=0}^{W-1} x_c(h, w) \quad (2)$$

Where z_c^H and z_c^W separately encode spatially-ordered context along the vertical and horizontal orientations for the c -th channel slice.

The assembled iEMA block conforms to an inverted residual topology laced with embedded attention. The forward traversal can be formalized as follows:

$$X_{out} = X_{in} + \text{Conv}_{1 \times 1} (\text{DWConv}_{3 \times 3} (\text{EMA} (\text{Conv}_{1 \times 1} (X_{in})))) \quad (3)$$

Where $\text{Conv}_{1 \times 1}(\cdot)$ undertakes pointwise convolution effecting channel-axis modulation, $\text{EMA}(\cdot)$ administers computationally lean spatial attention through paired 1D pooling operations, and $\text{DWConv}_{3 \times 3}(\cdot)$ gleans localized textural motifs via depthwise-separable convolution. The bypass linkage ensures fidelity of information preservation throughout the entire transformation chain.

RepViT-iEMA block

Computational thriftiness in the RepViT block stems from structural reparameterization: a multi-branch layout exemplified by RepVGGDW permits expressive representational learning during the training phase, while at deployment these parallel paths are consolidated into a singular convolution to maximise throughput. Figure 7a illustrates that the canonical RepViT block is organised around the token-mixer and channel-mixer design pattern.

$$X_{token} = \text{SE} (\text{RepVGGDW} (X_{in})) \quad (4)$$

$$X_{channel} = \text{Conv}_{1 \times 1} (\text{GELU} (\text{Conv}_{1 \times 1} (X_{token}, 2C))) \quad (5)$$

$$X_{out} = X_{in} + X_{channel} \quad (6)$$

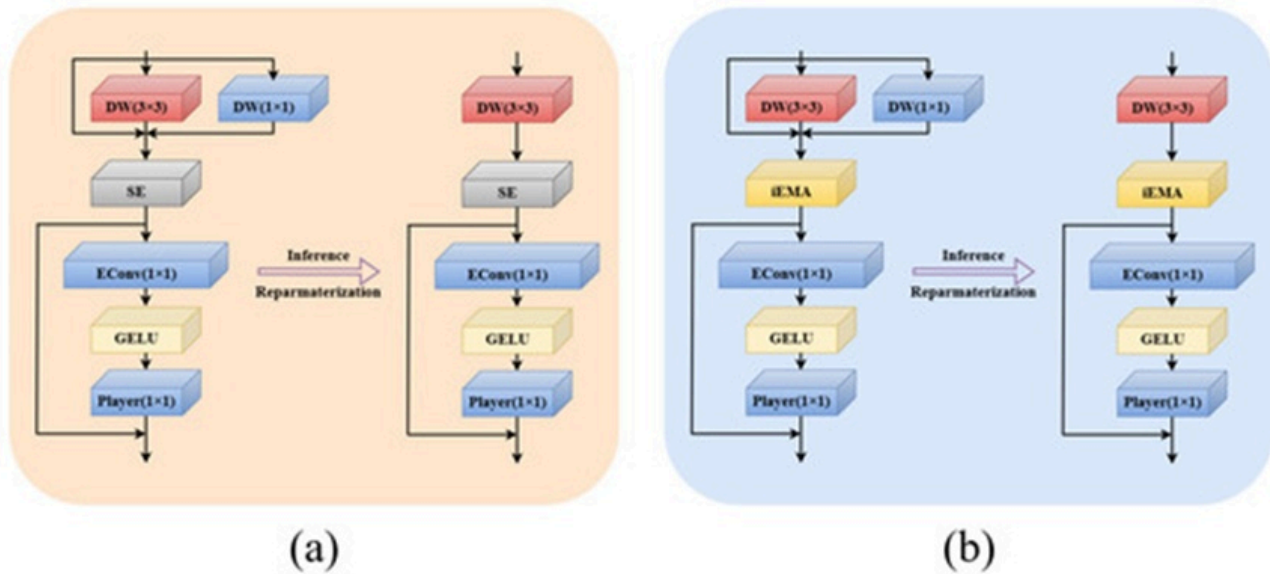


Figure 7. Reparameterization process diagram: (a) RepViT block and (b) RepViT-iEMA block.

That said, the SE attention operator embedded in the RepViT block remains blind to positional coding across the two spatial axes. This blindness curtails its capacity to register the faint textural shifts upon which fish maw genuineness

adjudication pivots. To rectify this gap, we discard SE attention in favour of our custom iEMA module, producing the RepViT-iEMA block as drawn in **Figure 7b**:

$$X_{token} = \text{iEMA} (\text{RepVGGDW} (X_{in})) \quad (7)$$

$$X_{channel} = \text{Conv}_{1 \times 1} (\text{GELU} (\text{Conv}_{1 \times 1} (X_{token}, 2C))) \quad (8)$$

$$X_{out} = X_{in} + X_{channel} \quad (9)$$

Where RepVGGDW(·) carries out reparameterizable depthwise convolution, GELU(·) references the Gaussian Error Linear Unit activation whose smooth, non-monotonic profile aids gradient flow, C captures the incoming channel dimensionality, and the multiplier 2C reflects channel expansion intrinsic to the inverted residual topology. This substitution keeps reparameterization benefits intact while sharpening spatial awareness via the cross-spatial learning pathways built into iEMA.

Dynamic adaptive multi-scale pyramid processing

Screening fish maw specimens demands multi-granularity handling that flexes according to diverse dimensions and contours. Veteran SPPF modules labour under frozen pooling schedules that shed high-frequency detail and offer no input-conditional adjustment; the upshot is impoverished representational quality.

We therefore engineer Dynamic Adaptive Multi-Scale Pyramid Processing (DAMSPP), shown in **Figure 8**, which adaptively modulates feature encoding through an Adaptive Multi-Scale Inception Block (AMSIB), depicted in **Figure 9**. Within AMSIB, Adaptive Inception Convolution teams with Dynamic Kernel Weighting to acquire scale-wise relevance scores, paired with shared convolutional weights for parsimony. The net result is adaptive textural capture across multiple resolutions without sacrificing fine detail.

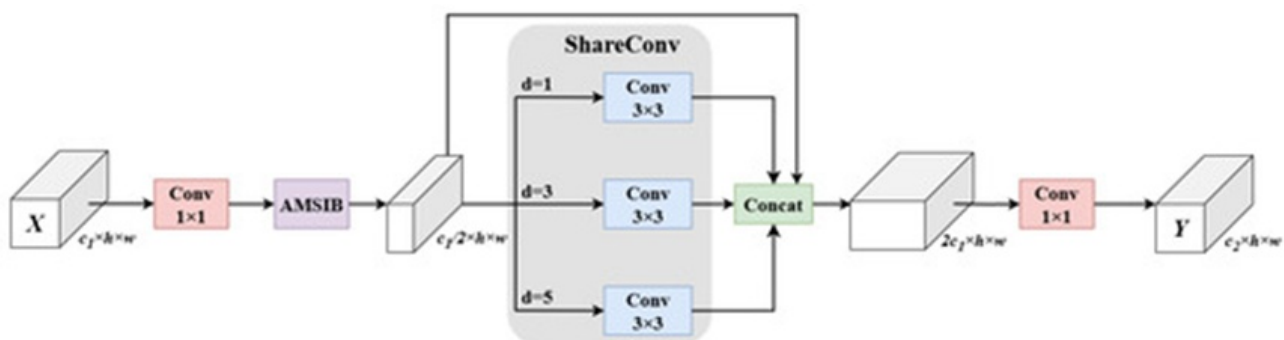


Figure 8. Structure of the DAMSPP module.

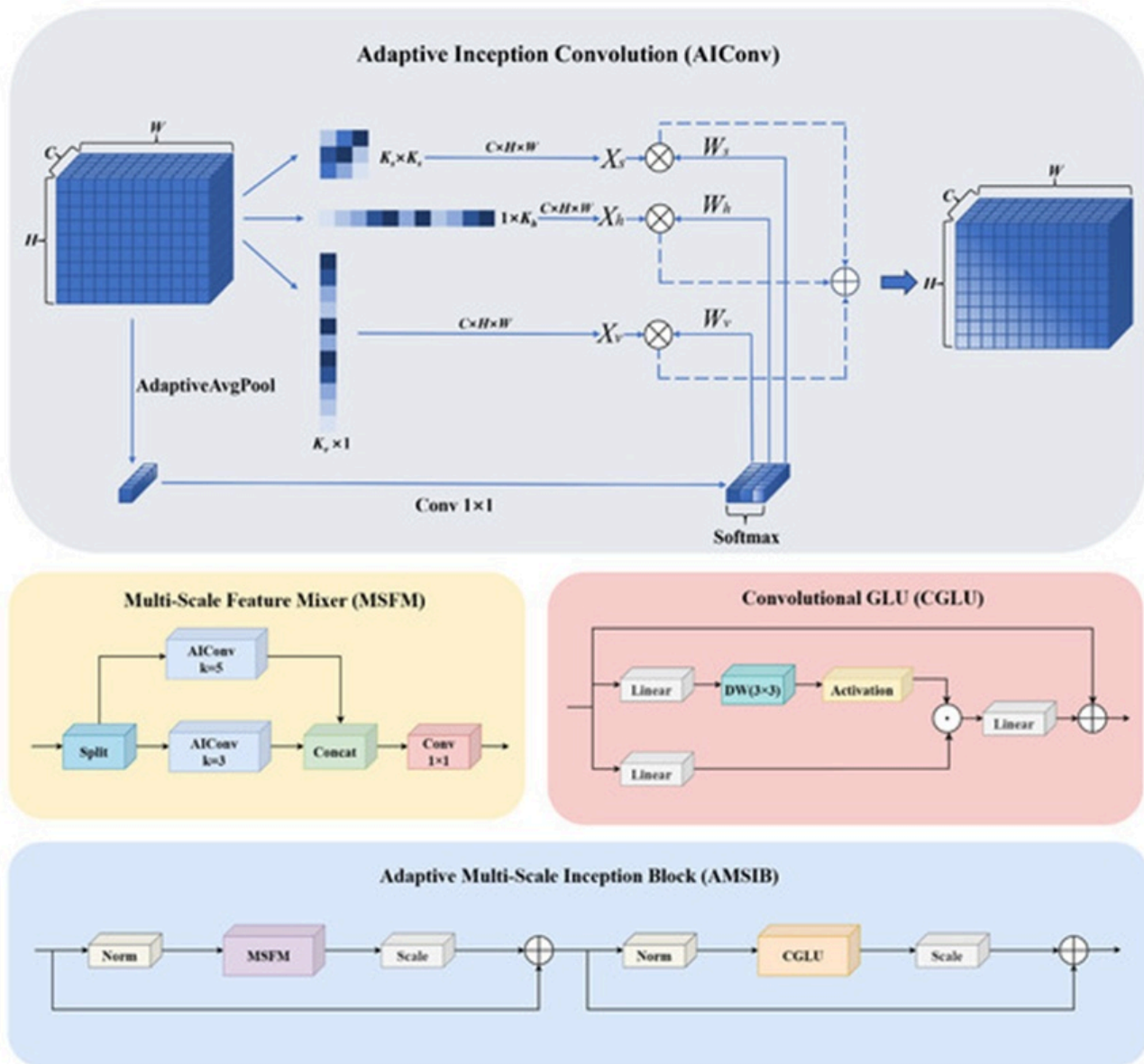


Figure 9. Structure of the AMSIB module.

Adaptive inception convolution

We begin by formulating Adaptive Inception Convolution (AIConv), whose linchpin is a Dynamic Kernel Weighting (DKW) routine that delivers globally adaptive processing of incoming feature maps.

Three parallel depthwise separable convolutional strands are instantiated for multi-scale extraction, respectively adopting square, horizontally oriented, and vertically oriented kernel configurations to harvest spatially directed signals. Multi-scale extraction is expressible as:

$$X_i = \text{DWConv}_i(X_m), \quad i \in \{s, h, v\} \quad (10)$$

Where X_{in} denotes the input tensor and $DWConv_s$, $DWConv_h$, $DWConv_v$ correspond to the square-kernel, horizontal-kernel, and vertical-kernel depthwise convolutions.

The kernel dimensions additionally abide by the inter-scale constraint:

$$k_h, k_v = 3 \times k_s + 2. \quad (11)$$

During weight synthesis, the DKW routine digests global summary statistics from the input and dynamically produces bespoke scaling factors for each processing strand. Its formulation is:

$$W = \phi(\text{Conv}_{1 \times 1}(\text{GAP}(X_m))) \quad (12)$$

Where $(\cdot)$ indicates the softmax normalisation and GAP stands for Global Average Pooling. This step first collapses spatial dimensions into a global descriptor, then projects the channel count from C to $3C$ via pointwise convolution so as to emit attentive weights for the three branches.

Each subtensor W_i acts as the weighting coefficient for its associated branch. The final AIConv output materialises through weighted aggregation:

$$\tilde{X} = \sum_{t \in \{s, h, v\}} W_t \odot X_t \quad (13)$$

$$X_{out} = \text{SiLU}(\text{BN}(\tilde{X})) \quad (14)$$

Where $\odot$ represents element-wise product, $\tilde{X}$ designates the dynamically fused map, and the blended features undergo batch normalisation followed by SiLU activation to promote stable training and swifter convergence.

Multi-scale feature mixer

To promote cross-resolution feature blending and informational synthesis, a Multi-Scale Feature Mixer (MSFM) grounded in AIConv is designed and implemented. Its mathematical description follows.

$$Y_i = \text{Split}(X_m), \quad i = 1, 2 \quad (15)$$

$$\hat{Y}_1 = \text{AIConv}_{3 \times 3}(Y_1) \quad (16)$$

$$\hat{Y}_2 = \text{AIConv}_{5 \times 5}(Y_2) \quad (17)$$

$$\tilde{Y} = \text{Conv}_{1 \times 1}(\text{Concat}(\hat{Y}_1, \hat{Y}_2)) \quad (18)$$

Operationally, the input feature map X_{in} is first split uniformly along the channel axis, furnishing two sub-bands Y_1 and Y_2 . These sub-bands are then separately fed to AIConv units tuned to differing receptive-field spans, yielding differentiated feature encodings. Channel-wise splicing followed by channel modulation ultimately yields the fused output Y' .

Adaptive multi-scale inception block

Finally, the MetaFormer design framework [36] is invoked to construct AMSIB, with MSFM serving as the token-mixing nucleus and the Convolutional Gated Linear Unit [37] borrowed from TransNeXt operating as the position-wise feedforward network. Trainable scaling coefficients are injected to realise precise feature blending regulation. The forward computation proceeds as:

$$Z_1 = X_m + \gamma_1 \cdot \text{MSFM}(\text{BN}(X_m)) \quad (19)$$

$$Z_2 = Z_1 + \gamma_2 \cdot \text{CGLU}(\text{BN}(Z_1)) \quad (20)$$

Where γ_1 and γ_2 are learnable modulation scalars seeded at 0.01 to maintain training equilibrium. Residual bypassing is embraced to secure faithful information propagation and well-behaved gradient backflow.

In summary, AMSIB applies content-dependent adaptive weighting across multi-resolution features, allowing faithful registration of delicate surface textures on fish maw specimens while preserving fine detail and amplifying overall representational strength.

Adaptive task-aligned detection head

Reliable fish maw screening hinges on harmonious cooperation between the classification and bounding-box regression sub-tasks, for correct species assignment presumes accurate spatial localisation. Stock YOLO11 detection heads employ a decoupled, two-branch architecture encumbered by a pair of fundamental flaws: first, their isolated processing paths obstruct the cross-task flow of information, giving rise to misalignment where confident class predictions co-occur with poor positional estimates; second, maintaining separate branches engenders duplicated feature extraction and outsized parameter consumption. Such weaknesses undercut both accuracy and runtime efficiency, particularly when confronting fish maw specimens of irregular contour against cluttered backdrops.

To surmount these hurdles, we draw inspiration from TOOD [38] and, as portrayed in **Figure 10**, introduce an Adaptive Task-Aligned Detection Head (ATADH). In the first instance, task-aligned assignment of training samples enforces congruity between the classification and localisation objectives. Secondly, a multi-layer shared-convolution schema featuring hierarchical feature aggregation trims parameter wastage while fostering effective informational exchange between tasks. Thirdly, the incorporation of a Task Decomposition unit augmented with dynamic kernel modulation alongside a DyDCNv2 operator realises clean separation of the two sub-tasks through tiered attention, while simultaneously enabling adaptive fitting of the non-uniform geometries typical of fish maw. This synergistic blueprint markedly shrinks the parameter count through task interaction, while efficiently disambiguating target coordinates, category, and confidence—directly tackling the difficulties of fish maw detection within complex visual environments.

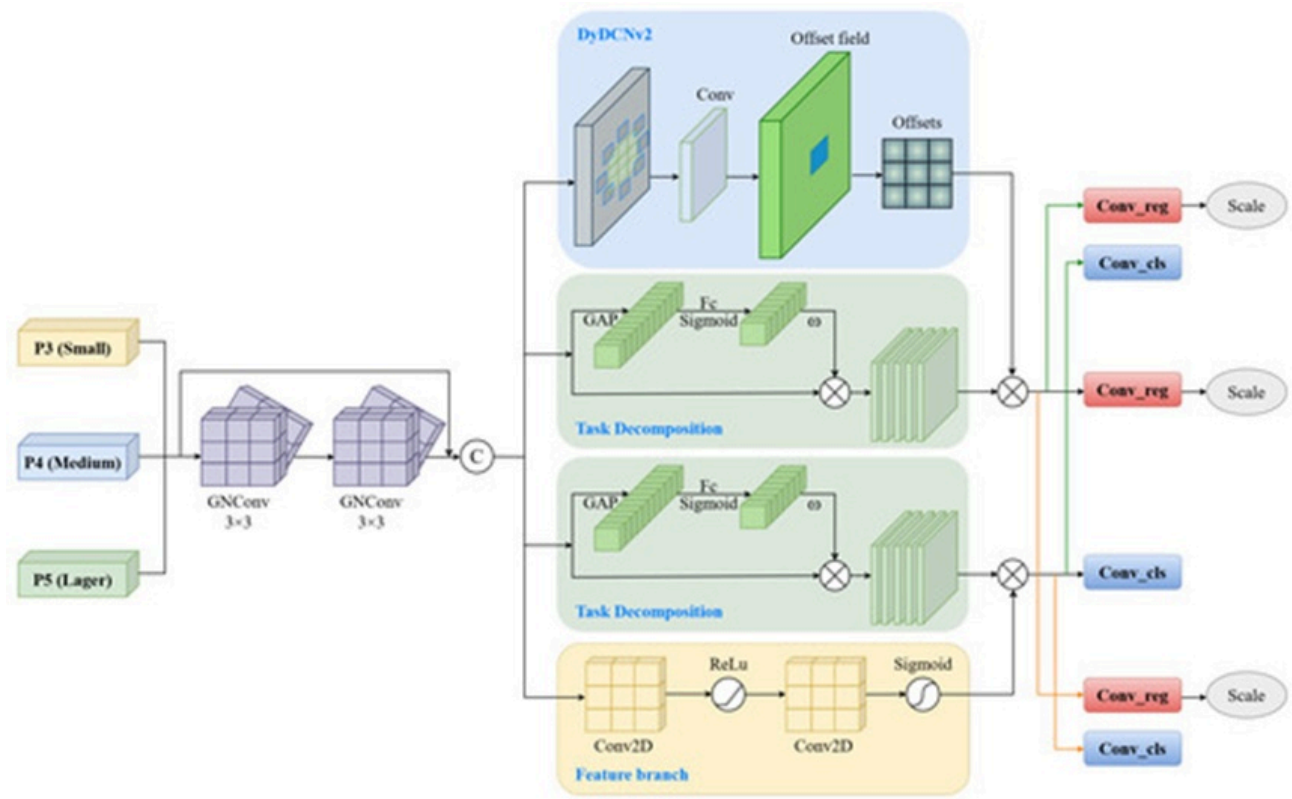


Figure 10. Structure of the ATADH module.

ATADH first commissions a multi-layer shared-convolution backbone for initial feature extraction, with hierarchical feature fusion lending augmented representational potency to the resulting descriptors. In practice, input feature maps are sequentially passed through two shared convolutional strata; the outputs emanating from each layer are subsequently spliced together and merged, forming rich, multi-depth feature representations. Such a scheme not only distils semantic information at several levels of abstraction but also facilitates gradient circulation, thereby improving training dynamics.

The Task Decomposition module stands as the operational core of ATADH, delivering clear separation and cooperative interaction between the classification and regression pipelines. This unit employs layered attention mechanisms to single out and intensify features pertinent to the task at hand, forestalling inter-task feature conflict. Its computational mechanics are captured as:

$$f' = \text{GAP}(f) \quad (21)$$

$$W = \sigma(\text{Conv}_{1 \times 1}(\text{ReLU}(\text{Conv}_{1 \times 1}(f')))) \quad (22)$$

$$\mathbf{K}_{\text{dynamic}} = W \odot \mathbf{K}_{\text{base}} \quad (23)$$

$$f_{\text{task}} = \text{ReLU}(\text{GN}(\text{Conv}_{\text{dynamic}}(f, \mathbf{K}_{\text{dynamic}}))) \quad (24)$$

Where f denotes the input feature map, GAP abbreviates Global Average Pooling, f' indicates the pooled feature vector, $\sigma(\cdot)$ is the sigmoid activation, W signifies the dynamically generated hierarchical attention weight tensor, K_{base} corresponds to the learnable base kernel parameters, $K_{dynamic}$ represents the dynamically conditioned kernel, $GN(\cdot)$ marks Group Normalisation, and $\odot$ stands for tensor broadcast multiplication.

The input f is first reduced via global average pooling to a compact descriptor f' , from which attention weights W are synthesised through a pair of convolutional projections; these weights W subsequently modulate the convolutional kernel, transforming the original feature f to yield task-specialised representations. Within this task-decomposition framework, the classification branch and regression branch function independently to achieve explicit task separation: the former is devoted to semantic cue extraction, the latter to geometric characterisation.

The DyDCNv2 unit exploits features from intermediate layers to dynamically infer spatial displacements and modulation masks, steering convolution such that it conforms to the erratic boundaries and textural motifs of fish maw specimens. Its defining relations are:

$$\{\Delta x, \Delta y, M\} = \text{Conv}_{\text{spatial}}(f) \quad (25)$$

$$M = \sigma(M) \quad (26)$$

$$y_{i,j}^{(\text{DyDCNv2})} = \sum_{m,n} x_{i+m+\Delta x_{i,j,m,n}, j+n+\Delta y_{i,j,m,n}} \cdot W_{m,n} \cdot M_{i,j,m,n} \quad (27)$$

Where M stands for the modulation mask, while Δx and Δy refer to the learnt horizontal and vertical offset grids, respectively; additionally, $\text{Conv}_{\text{spatial}}$ denotes a dedicated convolutional operator responsible for producing offsets and masks. Within the deformable convolution step, W embodies the kernel weight tensor, with i, j enumerating output spatial positions and m, n indexing relative coordinates within the sampling window. The sigmoid function is represented by $(\cdot)$.

The input feature f feeds a spatial convolution that concurrently emits displacement fields and masks, the latter being gated through the sigmoid nonlinearity. Dynamic deformable convolution then displaces the sampling grid in lockstep with the learnt offsets and reweights response magnitudes according to the mask coefficients, thereby accomplishing adaptive encoding of irregular geometries.

In closing, ATADH squarely addresses the core obstacles of fish maw genuineness verification—including background clutter interference, variability in target dimensions, and non-uniform texture topologies—by fusing dynamic feature adaptation, a lightweight design philosophy, and adaptive task decomposition. This integrated technical framework substantially boosts detection accuracy and system dependability for fish maw authenticity assessment under challenging real-world conditions.

Optimized loss function

The YOLO11 pipeline adopts CIoU as its bounding-box regression objective for fish maw screening. This metric jointly accounts for the overlap fraction, the Euclidean gap between centres, and the dimensional agreement linking predicted boxes to their annotated counterparts, as codified in [Eqs. 28–30](#):

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b_{gt})}{c^2} + \alpha v \quad (28)$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w_{gt}}{h_{gt}} - \arctan \frac{w}{h} \right)^2 \quad (29)$$

$$\alpha = \frac{v}{1 - IoU + v} \quad (30)$$

Where IoU measures the volumetric intersection ratio of reference and predicted detection zones, $\rho^2(b, b_{gt})$ captures the squared centre-to-centre Euclidean displacement, c corresponds to the diagonal length of the tightest enclosing rectangle, α functions as an adaptive balancing term, v registers aspect-ratio fidelity, w and h denote the predicted box extents, while w_{gt} and h_{gt} signify the ground-truth width and height.

A closer examination of [Eqs. 29 and 30](#) uncovers a pivotal flaw: when the moment aspect ratios coincide, the corrective factors α and v become entirely inert. Prevailing regression-optimisation scholarship leans heavily on geometric correspondence, systematically underplaying the shape diversity and size range that prove essential for fish maw detection targets.

For categories marked by dramatic morphological spread—fish maw being a case in point—Shape-IoU [\[39\]](#) brings object contour properties alongside scale parameters directly into the regression cost formulation, as laid out in [Eqs. 31–36](#):

$$L_{Shape-IoU} = 1 - IoU + distance^{shape} + 0.5 \times \Omega^{shape} \quad (31)$$

$$w_w = \frac{2 \times (w_{gt})^{scale}}{(w_{gt})^{scale} + (h_{gt})^{scale}} \quad (32)$$

$$h_h = \frac{2 \times (h_{gt})^{scale}}{(w_{gt})^{scale} + (h_{gt})^{scale}} \quad (33)$$

$$distance^{shape} = \frac{h_h \times (x - x_{gt})^2 + w_w \times (y - y_{gt})^2}{c^2} \quad (34)$$

$$\Omega^{shape} = \sum_{t=wh} (1 - e^{-\omega_t})^\theta, \quad \theta = 4 \quad (35)$$

$$\begin{cases} \omega_w = h_h \times \frac{|w - w_{gt}|}{\max(w, w_{gt})} \\ \omega_h = w_w \times \frac{|h - h_{gt}|}{\max(h, h_{gt})} \end{cases} \quad (36)$$

Where *distanceshape* encapsulates shape-sensitive distance penalty, *scale* conveys the scaling term, Ω^{shape} expresses the shape-based cost component, ω_w and h_h act as horizontal and vertical axis weightings, and x, y, x_{gt} , together with y_{gt} indicate centroid positions for predictions and ground truth, respectively.

Despite offering tangible gains relative to CloU, Shape-IoU does not adequately handle the degraded-quality exemplars that appear routinely in operational fish maw screening contexts. Field-deployed detection is beset by quality-eroding phenomena—shifts in camera angle, inconsistent illumination intensity, and uneven specimen curing degrees—that together inject marked quality variance into the assembled data and consequently deliver unbalanced gradient magnitudes during parameter updates.

Motivated by the Wise-IoU V3 formulation [40], the newly proposed Wise-ShapeloU objective marries the focusing coefficient r with Shape-IoU to enact quality-conditional gradient redistribution across anchor boxes. This design attenuates harmful gradient signals originating from poor-quality instances yet retains strong optimisation drive, as formalised in Eqs. 37–39:

$$L_{Wise-ShapeloU} = L_{WiseIoU} + distance^{shape} + 0.5 \times \Omega \quad (37)$$

$$L_{WiseIoU} = L_{IoU} \cdot r, \quad r = \frac{\beta}{\delta \cdot \alpha^{\beta-\delta}} \quad (38)$$

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \in [0, +\infty) \quad (39)$$

Where r stands for the non-monotonic focusing term, β serves as the outlier indicator, α and δ are tunable hyperparameters, and L_{IoU}^* represents the gradient amplification magnitude.

The focusing coefficient machinery accords elevated importance to moderately-scoring anchor boxes while gently diminishing the influence of top-tier anchors and substantially quashing disruptive gradient contributions arising from inferior samples. This quality-informed optimisation strategy directly confronts the environmental fluctuations and imaging inconsistencies inherent to fish maw detection, ultimately delivering elevated detection fidelity through more trustworthy quality-weighted anchor assignment.

Results and Discussion

Experimental settings and implementation details

The testing infrastructure comprised a Windows 11 environment, a 12th-Generation Intel® Core™ i5-12400F processor operating at 2.50 GHz, an NVIDIA GeForce RTX 4060 graphics unit, and the PyTorch 1.12.1 deep-learning toolkit with CUDA 11.3 acceleration. Training hyperparameters were prescribed as follows: a 200-epoch training budget, mini-batch cardinality of 16, a starting learning rate of 0.01, no reliance on pre-trained weight parameters, the SGD optimiser instantiated with a momentum coefficient of 0.937 and weight decay strength of 0.0005, and input imagery normalised to a 640×640 pixel grid.

Evaluation metrics

For a holistic assessment of model behaviour, the present work drew on a suite of fundamental performance yardsticks. Chief among these, mean Average Precision (mAP) was tallied under two Intersection-over-Union threshold conventions, designated mAP@0.5 and mAP@0.5:0.95. Precision (P) tracks the proportion of correctly asserted positive predictions relative to the full set of positive classifications emitted by the model, whereas Recall (R) reflects the model's aptitude for surfacing all positive exemplars contained in the ground-truth corpus. Average Precision (AP) computes the integrated area beneath the precision–recall trajectory. The 0.5 IoU setting aggregates AP readings across every object class, whereas mAP@0.5:0.95 reports the mean of AP values evaluated over an array of IoU cutoffs stepping from 0.5 through 0.95 using 0.05 increments. Their mathematical formulations are given below:

$$P = \frac{TP}{TP + FP} \quad (40)$$

$$R = \frac{TP}{TP + FN} \quad (41)$$

$$AP_I = \int_0^1 P_I(R_I) dR_I \quad (42)$$

$$mAP = \frac{1}{n} \sum_{I=1}^n AP_I \quad (43)$$

Where n conveys the overall number of classes, TP enumerates positive instances correctly retrieved, FP counts negative instances erroneously promoted as positive, and FN tallies positive instances that escaped model retrieval.

Comparison experiment

Benchmark algorithm of prior experiments

To pinpoint the YOLO-generation variant most favourable for fish maw genuineness verification, an extensive side-by-side evaluation of assorted YOLO models was first carried out. The comparison canvassed mAP@0.5, mAP@0.5:0.95, parameter count, floating-point operation volume, and storage occupancy; the aggregated findings are collated in [Table 2](#). The head-to-head assessment showed that while YOLOv8n attained the strongest raw accuracy at 90.6% mAP@0.5, its parametric overhead, computational demand, and on-disk size were comparatively burdensome. In marked distinction,

YOLO11n delivered a highly advantageous balance across all axes, reaching 90.5% mAP@0.5 and 84.9% mAP@0.5:0.95 whilst occupying a storage envelope of merely 5.2 MB and carrying only 2.58 M parameters. This performance-to-footprint equilibrium establishes YOLO11n as an exceptionally fitting backbone for fish maw authenticity screening under hardware-limited deployment conditions.

Table 2. Benchmark algorithm of the prior experiment.

Model	mAP@0.5/%	mAP@0.5:0.95/%	Param/M	FLOPs/G	Size/MB
YOLOv8n	90.6	84.4	3.01	8.1	6
YOLOv9t	89.5	84.3	2.8	11.7	23.4
YOLOv10n	90	84.7	2.7	8.2	5.5
YOLO11n	90.5	84.9	2.58	6.3	5.2
YOLOv12n	89.2	83.1	2.51	5.8	5.2

Comparison of different SPPF

To determine the relative merit of the DAMSPP element engineered within this investigation, a battery of controlled comparisons was devised. Three alternative designs were measured against the baseline SPPF, namely SPPELAN [41], AIFI [42], and the proposed DAMSPP; the consolidated evidence is presented in Table 3.

Table 3. Comparison of different SPPF module improvements.

Model	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	Param/M	Size/MB
SPPF	85.7	81.8	90.5	84.9	2.58	5.2
SPPELAN	88.7	78.4	88.3	82.5	2.43	4.9
AIFI	87.1	84.2	91.5	85.7	3.21	6.4
DAMSPP	88.8	85.4	92.2	86.7	2.81	5.7

The collected findings indicate that although adopting the DAMSPP module entailed a marginal expansion in computational burden and parameter footprint, the system-level gains proved substantial. Across the four principal diagnostic metrics, DAMSPP recorded 88.8% precision, 85.4% recall, 92.2% mAP@0.5, and 86.7% mAP@0.5:0.95. Taken together, the empirical record affirms the soundness of the DAMSPP architectural concept advanced herein.

Comparison of different detection heads

Extensive comparative testing was conducted to appraise the effectiveness of the ATADH detection head formulated as part of this research. Four detection-head variants—DyHead [43], Aux [44], MultiSEAMHead [45], and ATADH—were set against the factory YOLO11 detection head, with results catalogued in Table 4.

Table 4. Comparison of different detection head modules.

Model	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	Param/M	Size/MB
11-Head	85.7	81.8	90.5	84.9	2.58	5.2
DyHead	88	81.2	89.7	85	3.1	6.3

Aux	85.5	82.1	90.5	85.1	2.31	4.7
MultiSEAMHead	87.4	80.6	91.8	85.7	4.6	9.2
ATADH	88.6	86.4	92.1	87	2.2	4.5

The empirical data reveal that fitting the ATADH detection head delivered the strongest all-around readings on every monitored indicator. Regarding discriminative accuracy, ATADH posts 88.6% precision, 86.4% recall, 92.1% mAP@0.5, and 87% mAP@0.5:0.95. From a resource-efficiency standpoint, ATADH occupies a slender parameter budget of only 2.2 M alongside a 4.5 MB storage requirement, thereby realising a genuinely compact architectural footprint. The ATADH detection head thus demonstrates peerless efficacy when appraised through composite performance benchmarks.

Comparison of different convolution modules

To verify the added value of DyDCNv2 atop ATADH, four alternative configurations were examined under matched conditions: plain convolution, depthwise separable convolution, DCNv2, and DyDCNv2; the evaluation summary appears in [Table 5](#). The trial outcomes demonstrate that DyDCNv2 hits mAP@0.5 and mAP@0.5:0.95 marks of 92.1% and 87%, respectively, while incurring a near-identical parameter count. Set against plain convolution, these tallies reflect improvements of 1.5 and 1.9 percentage points, compellingly substantiating the potency of the DyDCNv2 dynamic deformable convolution strategy.

Table 5. Comparison of different convolution schemes in ATADH.

Model	Convolution	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	Param/M	FLOPs/G	Size/MB
ATADH	Conv	88.5	82.1	90.6	85.1	2.17	8	4.4
	DWConv	89.6	79.1	89.8	84.2	2.14	7.4	4.4
	DCNv2	81.9	82.5	90.4	83.8	2.18	7.6	4.4
	DyDCNv2	88.6	86.4	92.1	87	2.2	7.9	4.5

Comparison of different loss functions

To ascertain the benefits of the newly introduced loss formulation, head-to-head experiments were staged between Wise-ShapeloU and widely adopted alternatives—CloU, DIoU [\[46\]](#), GIoU [\[47\]](#), EIoU [\[48\]](#), and PIoU [\[49\]](#)—with the resulting figures collected in [Table 6](#) and the loss descent trajectories during model fitting charted in [Figure 11](#).

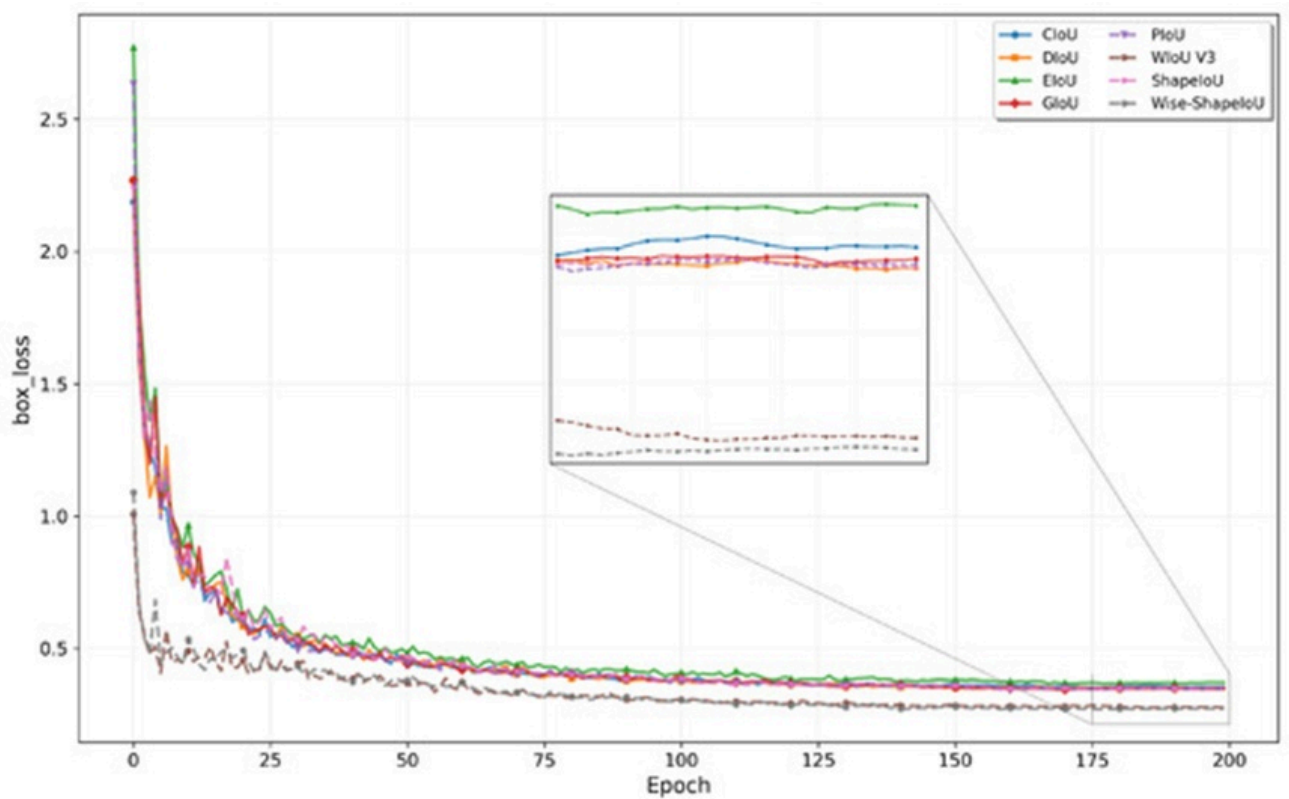


Figure 11. Loss function variation curves during training.

Table 6. Comparison of different loss functions.

Method	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%
CIoU	89.4	88.1	93.4	88.5
DIoU	87.9	89.2	93.9	87.6
GIoU	88.1	88.6	93.5	87.3
EIoU	91.5	84.6	93.3	87.5
PIoU	90.7	86.5	93.4	88.3
WIoU V3	91.6	86.9	93.7	88.7
ShapeloU	90.9	86.1	93.7	88.8
Wise-ShapeloU	91.9	89.6	94	88.9

Cross-comparison reveals that WIoU V3 and ShapeloU each display enhanced behaviour both in optimisation convergence and in terminal performance readings. Wise-ShapeloU selectively amalgamates the complementary advantages of the WIoU V3 and ShapeloU criteria, manifesting smoother convergence dynamics during the training cycle and clinching top-tier results on all primary performance gauges. In specific terms, Wise-ShapeloU achieves 91.9% precision, 89.6% recall, 94% mAP@0.5, and 88.9% mAP@0.5:0.95, thereby yielding the most favourable detection performance observed.

Comparison of RDT-YOLO with mainstream models

To rigorously probe the capacity of the proposed RDT-YOLO framework for fish maw genuineness classification, controlled benchmarking was orchestrated against a suite of prominent contemporary models, with experimental conditions and data splits held fixed to guarantee impartial measurement. The exhaustive findings are consolidated in **Table 7**. Although YOLOv12s attained an mAP@0.5:0.95 figure of 89.3%—edging past RDT-YOLO by 0.4 percentage points—its parameter volume, floating-point throughput, and on-disk footprint all substantially outstripped those of RDT-YOLO. Moreover, RDT-YOLO eclipsed rival mainstream architectures on all four cardinal detection indices, registering 91.9% precision, 89.6% recall, 94% mAP@0.5, and 88.9% mAP@0.5:0.95, corresponding to absolute lifts of 6.2, 7.8, 3.5, and 4 percentage points vis-à-vis the YOLO11n reference model.

Table 7. Comparison of different mainstream models.

Model	Backbone	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	Param/M	FLOPs/G	Size/MB
Faster-RCNN	ResNet50	77.2	75.8	85.7	71.5	41.37	134	377
Cascade-RCNN	ResNet50	88.9	78.3	88.2	72.6	69.21	162	484
SSD	VGG16	74.7	84.7	83.8	66.2	24.28	61.15	92.11
TOOD	ResNet50	83.9	63.3	80.3	73.7	32.03	123	294
DINO	ResNet50	81.4	69.6	84	79.2	47.55	179	374
RTDETR	ResNet18	89.8	80.1	83.3	77.5	19.88	57	38.6
YOLOv10s	—	90.7	87.2	93.2	89.1	8.04	24.5	15.8
YOLO11n	—	85.7	81.8	90.5	84.9	2.58	6.3	5.2
YOLO11s	—	86.8	87	93.5	88.8	9.41	21.3	18.3
YOLOv12s	—	87.8	83.7	93.5	89.3	9.08	19.3	17.8
RDT-YOLO(Ours)	—	91.9	89.6	94	88.9	2.3	7.7	4.8

Measured alongside two-stage pipelines, specifically Faster R-CNN and Cascade R-CNN, RDT-YOLO posts mAP@0.5 advantages of 8.3 and 5.8 percentage points. When contrasted with single-stage contenders TOOD and SSD, the mAP@0.5 margins widen to 10.2 and 13.7 percentage points, respectively. Relative to the transformer-based detectors DINO and RT-DETR, RDT-YOLO records mAP@0.5 enhancements of 10 and 10.7 percentage points. In comparison with peer single-stage models YOLOv10s and YOLOv12s, RDT-YOLO secures mAP@0.5 leads of 0.8 and 0.5 percentage points. Notably, the mAP@0.5 score achieved by RDT-YOLO surpasses that of YOLO11s—a bulkier sibling within the same model series—by 0.5 percentage points.

Concerning computational expenditure, although RDT-YOLO consumes 1.4 GFLOPs more than the bare YOLO11n backbone, it records the leanest FLOPs, parameter tally, and model size among all other evaluated systems. Considered holistically, RDT-YOLO captures the premier overall standing within the comparator set while furnishing superior detection accuracy; concurrently, it displays outstanding frugality in both model heft and inferential speed. These attributes render it especially well-suited to latency-critical screening contexts, a property of direct practical relevance to operational fish maw authenticity verification.

Ablation experiment

Positional ablation study of C3k2-RVB and C3k2-RVB-iEMA modules

To judge the efficacy of the tiered placement scheme, controlled evaluations were performed as enumerated in **Table 8**. In this tabulation, Base denotes the stock C3k2 module arrangement as the control configuration; Strategies A and B

deploy uniform replacement, substituting every backbone C3k2 module with C3k2-RVB and C3k2-RVB-iEMA modules, respectively; lastly, Strategy C institutes the layered hybrid approach, wherein the initial two shallow C3k2 modules give way to C3k2-RVB modules for fortified low-level feature encoding, while the terminal two deep C3k2 modules are displaced by C3k2-RVB-iEMA modules for augmented high-level semantic abstraction.

Table 8. Experiments with different positions of C3k2-RVB and C3k2-RVB-iEMA modules.

Model	Size/MB	FLOPs/G	Param/M	mAP@0.5:0.95/%	mAP@0.5/%	R/%	P/%
Base	5.2	6.3	2.58	84.9	90.5	81.8	85.7
A	5	6	2.44	84.8	90.7	81.6	87.7
B	5.1	6.1	2.45	85.1	90.7	83.4	82.9
C	5.1	6.1	2.45	86.4	91.9	85.6	85.6

Strategy C produces the optimal outcomes on both mAP@0.5 and mAP@0.5:0.95 readouts, outperforming Base, Strategy A, and Strategy B while sustaining a broadly analogous level of structural complexity. These observations verify that Strategy C aligns more effectively with the operational requirements of fine-grained textural discrimination in fish maw inspection.

RDT-YOLO model overall ablation study

A systematic ablation exercise was undertaken to holistically gauge the contribution of each proposed enhancement atop the YOLO11n foundation, with evaluations conducted on the held-out test set. Every experimental run shared an identical environment configuration and hyperparameter recipe. The labels Base, A, B, C, and D refer respectively to the unaltered baseline, the C3k2-RVB plus C3k2-RVB-iEMA pairing, the DAMSPP unit, the ATADH unit, and the Wise-ShapeloU regression objective. The set of comparison criteria comprised six indicators: P, R, mAP@0.5, mAP@0.5:0.95, parameter volume (Param), and on-disk size (Size). The full ablation ledger is presented in **Table 9**.

Table 9. Ablation study results of the proposed RDT-YOLO model.

Base	+A	+B	+C	+D	Size/MB	Param/M	mAP@0.5:0.95/%	mAP@0.5/%	R/%	P/%
✓	-	-	-	-	5.2	2.58	84.9	90.5	81.8	85.7
✓	✓	-	-	-	5.1	2.45	84.9	91.8	85.6	85.5
✓	-	✓	-	-	5.7	2.81	86.7	92.2	85.4	88.8
✓	-	-	✓	-	4.5	2.2	87	92.1	86.4	88.6
✓	-	-	-	✓	5.2	2.58	86.3	92	82.6	91.2
✓	✓	✓	-	-	5.5	2.68	86.8	92.4	84.7	91.1
✓	✓	-	✓	-	4.3	2.07	86.7	92.1	87.6	89.8
✓	-	✓	✓	-	4.9	2.43	87.6	93.2	86.3	87.9
✓	✓	✓	✓	-	4.8	2.3	88.5	93.4	88.1	89.4
✓	✓	✓	✓	✓	4.8	2.3	88.9	94	89.6	91.9

Note: ✓ in the table indicates that the module is added.

Inspecting **Table 9** with YOLO11n as the anchor, it becomes evident that each architectural modification, when deployed in isolation, confers measurable detection accuracy dividends. When all four innovations are activated concurrently, the resultant mAP@0.5 surpasses that attainable through any single-module addition.

More concretely, grafting the refined C3k2-RVB and C3k2-RVB-iEMA modules onto YOLO11n enables the network to mine richer stratified information, pushing mAP@0.5 to 91.8% while concurrently compressing both the parameter budget and the storage envelope. Introducing the DAMSPP module thereafter lifts P to 91.1%, as both principal detection gauges rise markedly to 92.4% and 86.8%, thereby corroborating the value contributed by the DAMSPP module's adaptive feature conditioning and hierarchical extraction mechanics. Appending the custom-built ATADH detection head further amplifies cross-task interplay; the recorded figures exhibit comprehensive metric uplift, with the two detection accuracy readouts climbing to 93.4% and 88.5%, respectively, as the parameter tally shrinks to 2.3 M and the model footprint contracts to 4.8 MB. Finally, the inclusion of the Wise-ShapeloU criterion nudges detection precision to 94%, with P and R culminating at 91.9% and 89.6%.

Measured against the baseline architecture, the aggregate net gains amount to: a 6.2% rise in P, a 7.8% rise in R, a 3.5% lift in mAP@0.5, and a 4% uptick in mAP@0.5:0.95, accompanied by an 11% trimming of parameter count and a 7.7% shrinkage of storage volume. These results lend weight to the conclusion that the enhanced methodology materially advances fish maw authenticity discrimination fidelity whilst attaining computational leanness through slimmer parametrisation and a lighter model footprint.

Visual analytics

To render the merits of the augmented RDT-YOLO model more immediately perceptible, **Figure 12** juxtaposes visual outputs produced by YOLOv10s, YOLO11n, YOLO11s, YOLOv12s, and RDT-YOLO across detection contexts spanning a spectrum of difficulty—uncluttered scenes, cluttered backdrops, occlusion-heavy settings, and arrangements featuring densely clustered small objects. As the figure plainly illustrates, standard YOLO instantiations are routinely afflicted by detection mistakes, background-triggered false alarms, fragility under occlusion, and a struggle to disentangle tightly packed targets; YOLO11n proves especially deficient within dense-object regimes. RDT-YOLO, by contrast, harnesses its refined multi-scale feature fusion pipeline and strengthened representational capacity to localise targets correctly in every tested scenario, competently negotiating severe occlusions and reliably isolating individual objects amid crowded configurations. The qualitative evidence decisively affirms RDT-YOLO's potency, demonstrating that the proposed methodology possesses heightened environmental resilience and immunity to disturbance, thereby measurably lowering the incidence of missed targets and erroneous positives.

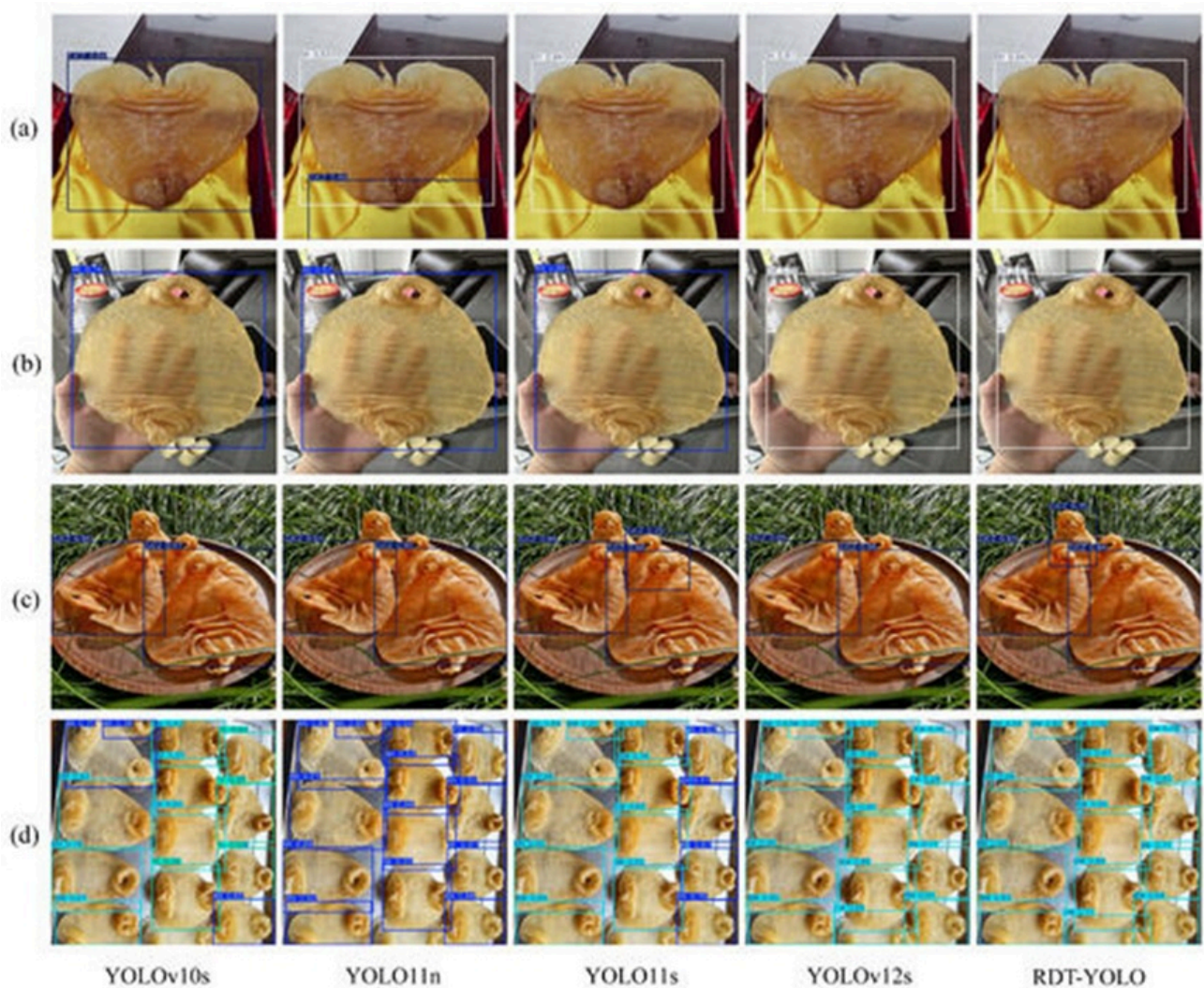


Figure 12. Comparison of test results for different models: (a) simple single-target detection scenario, (b) complex single-target detection with background interference, (c) occluded target detection with mutual occlusion, and (d) dense target detection with closely arranged targets.

To further illuminate the model's detection behaviour, three canonical scene categories were sampled from the test partition and subjected to heatmap rendering via KPCA-CAM, as displayed in [Figure 13](#). Scrutiny of these heatmap renderings discloses that RDT-YOLO concentrates its attentional activation with sharper precision across all settings compared with YOLO11n and YOLO11s. The refined model mitigates, to varying extents, the confounding influence of background noise upon target identification and exhibits more robust feature extraction when dealing with occlusion-challenged scenarios. Relative to the baseline algorithm, RDT-YOLO can autonomously acquire and zoom in on the textural signatures that hold diagnostic value. This substantially reduces background interference and the blurring of features among neighbouring targets, thus strengthening both the fidelity and the dependability of fish maw genuineness screening.

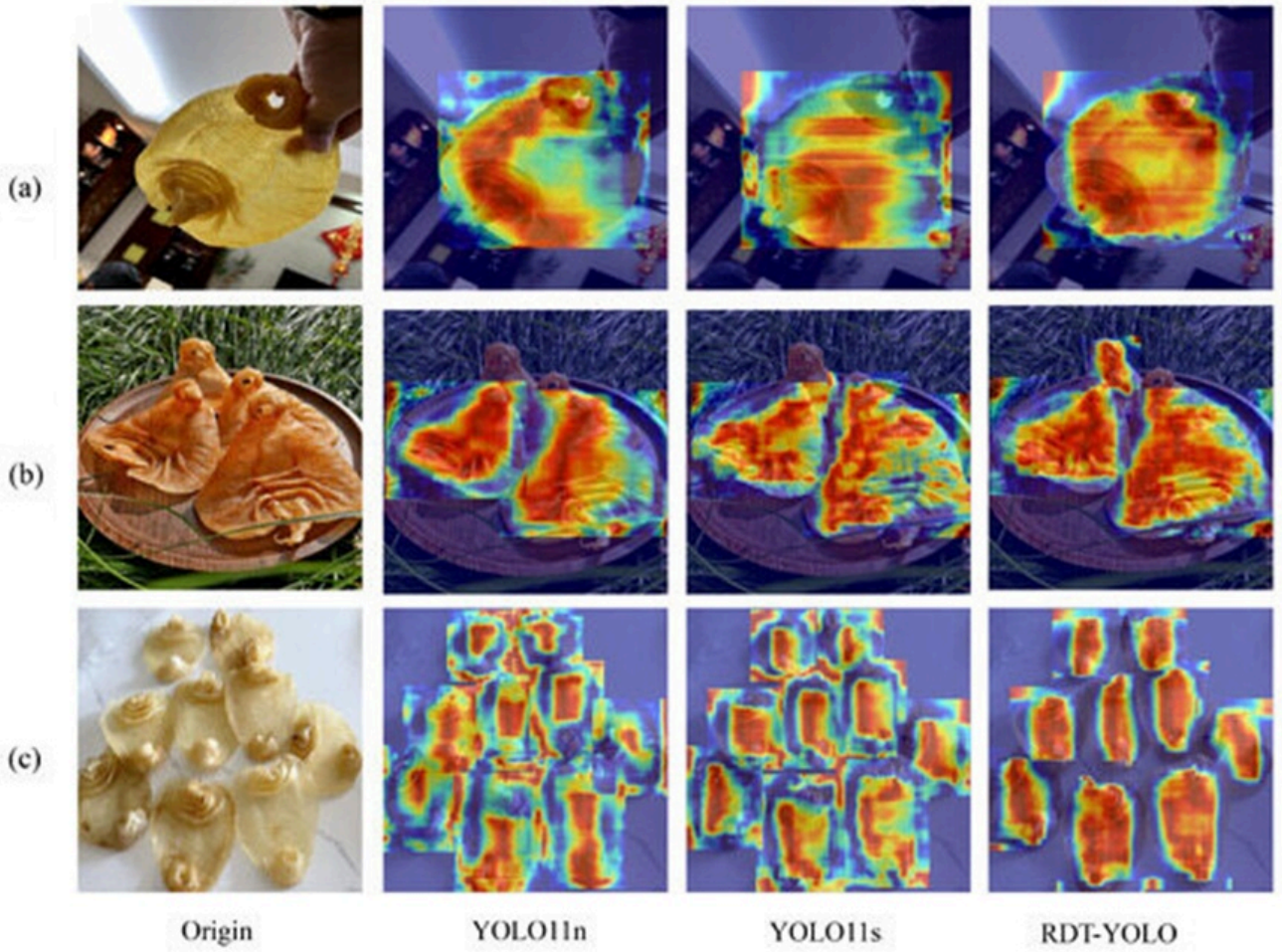


Figure 13. Comparison of heat maps: (a) single-target detection, (b) occluded target detection, and (c) dense target detection.

Generalization experiments

To interrogate the cross-domain portability of RDT-YOLO, the NEU-DET steel surface defect dataset [50] disseminated by Northeastern University was enlisted for out-of-domain testing. This dataset was deliberately chosen because steel defect inspection and fish maw authenticity verification impose comparable technical requirements—both demand exacting texture representation and the ability to resolve minute distinctions—thus constituting a meaningful proving ground for assessing generalisation capability. With the environmental configuration kept identical to that employed in the fish maw experiments, RDT-YOLO was benchmarked against leading object detection frameworks; the resultant data are collated in Table 10.

Table 10. Defect detection results on the NEU-DET dataset.

Model	Size/MB	FLOPs/G	Param/M	mAP@0.5:0.95/%	mAP@0.5/%	R/%	P/%
Faster-RCNN	377	134	41.37	38.9	76.1	69.5	72.4
Cascade-RCNN	484	162	69.21	41.4	76.6	69.9	72
SSD	92.11	61.15	24.28	35.2	71.2	68.2	71.7
TOOD	294	123	32.03	37.6	65.13	63.2	69.5

DINO	374	179	47.55	32.9	60.2	51.6	56.5
RTDETR	38.6	57	19.88	38	68.6	61.2	69.6
YOLOv10s	15.8	8.04	24.5	38	71.6	65.9	72.9
YOLO11n	5.2	6.3	2.58	40.9	75.4	69.3	69.6
YOLO11s	18.3	21.3	9.41	41.2	76.5	65.3	77.4
YOLOv12s	17.8	19.3	9.08	42.1	77.5	67.3	79.4
RDT-YOLO (Ours)	4.8	7.7	2.3	43	78.9	73.2	75.9

As **Table 10** attests, RDT-YOLO scores 78.9% mAP@0.5 and 43% mAP@0.5:0.95 on the NEU-DET corpus, outperforming every model in the comparison slate. Against the YOLO11n baseline, RDT-YOLO improves mAP@0.5 by 3.5 percentage points and mAP@0.5:0.95 by 2.1 percentage points, while simultaneously trimming parameter count and storage occupancy by 10.9% and 7.7%, respectively. When set alongside YOLOv10s, YOLO11s, and YOLOv12s, RDT-YOLO lands superior detection scores while retaining a decisive upper hand in parametric economy and computational lightness.

Figure 14 furnishes a side-by-side visual juxtaposition of YOLO11n and RDT-YOLO predictions on illustrative samples drawn from the NEU-DET repository. The topmost row reproduces the original frames, the middle row conveys YOLO11n's detection outcomes, and the lowermost row presents the outputs delivered by RDT-YOLO. One can clearly discern that YOLO11n is plagued by serious omission errors and category misattributions, whereas RDT-YOLO reliably pinpoints defect locations and correctly assigns defect classes, thus effectively redressing these failings.

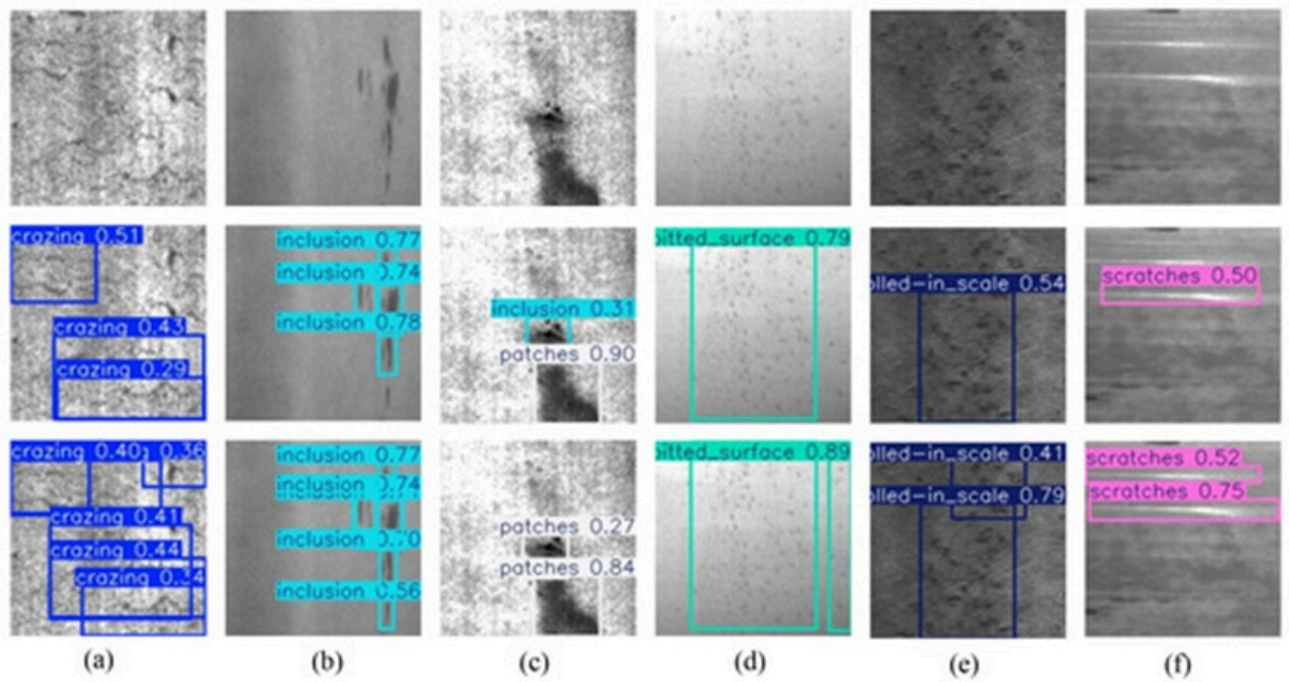


Figure 14. Visualization comparison experiments on the NEU-DET dataset: (a) crazing, (b) inclusion, (c) patches, (d) pitted surface, (e) rolled in scale, and (f) scratches.

The generalisation assay results robustly corroborate that RDT-YOLO is endowed with strong cross-domain transfer competence, demonstrating pertinence not merely to fish maw genuineness screening but also to industrial surface

anomaly detection and related tasks that hinge on fine-grained visual feature discrimination, thereby signalling expansive application potential.

Technical necessity and method advantages

Established protocols for verifying fish maw genuineness rest predominantly on seasoned human judgment and laboratory-based chemical assays. While these routes can yield dependable conclusions, they are beset by practical impediments: per-sample expenses run high, turnaround cycles stretch long, and both specialised hardware and trained analysts are indispensable. Such drawbacks leave them ill-matched to the operational realities of commercial environments, where screening must be high-volume, swift, and economical. As fish maw continues to appreciate in market value and fraudulent practices evolve in sophistication, the call for automated, accurate, and readily fieldable detection pipelines grows increasingly urgent.

Notwithstanding deep learning's proven track record across a spectrum of food-safety applications—spanning quality sorting of agricultural goods, spoilage assessment in meats, and leaf-grade classification in teas—academic enquiry concerning fish maw authenticity has stayed tethered to conventional analytical chemistry and spectroscopic traditions. The work reported here constitutes the first instance of transplanting deep-learning object detection into the arena of fish maw genuineness screening. The devised RDT-YOLO framework recasts the authentication exercise as a visual recognition problem, delivering a fully end-to-end automated workflow.

When measured against conventional techniques, the proposed model offers a constellation of clear merits: it dispenses with intricate sample preparation and manually crafted feature descriptors, instead learning to independently extract the visually discriminating hallmarks of fish maw; it boasts rapid inferential speed befitting real-time operation; its foundation within a deep-learning ecosystem affords considerable extensibility, easing subsequent model tuning and functional broadening; and it carries modest operational costs, as image acquisition relies solely on an ordinary digital camera or mobile phone, eschewing the need for costly specialist analytical hardware.

Limitations analysis and future research directions

Notwithstanding the strong outcomes RDT-YOLO delivers on fish maw genuineness tasks, certain restrictions persist. The dataset underpinning this study is still of finite breadth and drawn predominantly from a narrow set of vendors and geographical zones, which may introduce sampling skew. Fish maw specimens stemming from different source regions, subjected to distinct curing procedures, or held under varying storage regimes can differ subtly in external presentation, potentially undercutting the model's ability to generalise. Moreover, although the architecture achieves a notable degree of computational parsimony, further aggressive compression is still needed before deployment on edge-class devices with severely capped computing resources becomes straightforward. The repertoire of detectable categories remains relatively fixed as well, obliging periodic refreshing and enlargement of the model's class vocabulary as new modes of counterfeiting surface.

Trajectories for subsequent inquiry include: building substantially larger and more variegated fish maw corpora that incorporate specimens of diverse provenance, variety, qualitative tier, and known adulteration type; examining the utility of data augmentation and generative adversarial networks to counterbalance skewed class distributions; investigating multimodal fusion-based detection that marries visual surface cues with physicochemical signatures—embracing near-infrared and Raman spectral techniques—to erect a comprehensive and dependable recognition framework; and pushing forward with model miniaturisation through network pruning, knowledge distillation, quantisation, and analogous tactics to further boost inferential throughput.

Conclusion

Verifying the authenticity of fish maw is a cornerstone of regulatory enforcement and consumer defence within the fast-growing global fish maw sector. This contribution details the application of RDT-YOLO, grounded in computer vision, to the fish maw genuineness determination problem. By weaving together an inventive multi-modular design—embracing C3k2-RVB, C3k2-RVB-iEMA, and DAMSPP components coupled with an ATADH detection head—RDT-YOLO realises a finely tuned sweet spot between discriminative power and computational economy. The resultant model delivers a standout 94% mAP@0.5 yet remains remarkably lean, occupying just 4.8 MB of storage space. Its double-barrelled superiority over legacy detection approaches—in both accuracy and parameter parsimony—paves the way for fieldable, real-time use within live commercial settings. While this study supplies a scalable and resource-savvy pathway for fish maw inspection, further investigative steps are necessary to elevate the model's capacity to generalise. Augmenting the dataset, enveloping a richer range of market conditions and fish maw varieties, and stretching the algorithm's reach into additional food-safety screening tasks could all serve to fortify detection robustness.

Acknowledgements

None

Conflict of interest

None

Financial support

This work was supported by the National Natural Science Foundation of China (Grant No. 62373390), the Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2023A1515011230 and 2022B1515120059), and the Science and Technology Planning Project of Guangzhou (Grant No. 2023E04J1238).

Ethics statement

None

Received: 15 Sep 2021

Revised: 14 Oct 2021

Accepted: 26 Nov 2021

Published online: 15 March 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

FAO. The State of World Fisheries and Aquaculture 2024—Blue Transformation in Action. Rome: Food and Agriculture Organization of the United Nations; 2024.

Ben-Hasan A, de Mitcheson YS, Cisneros-Mata MA, Jimenez EA, Daliri M, Cisneros-Montemayor AM, et al. China's fish maw demand and its implications for fisheries in source countries. *Mar Policy*. 2021;132:104696.

Chen Y, Jin H, Yang F, Jin S, Liu C, Zhang L, et al. Physicochemical, antioxidant properties of giant croaker (*Nibea japonica*) swim bladders collagen and wound healing evaluation. *Int J Biol Macromol*. 2019;138:483-91.

Dai C, Dai L, Yu FJ, Li XN, Wang GX, Chen J, et al. Chemical and biological characteristics of hydrolysate of crucian carp swim bladder: Focus on preventing ulcerative colitis. *J Funct Foods*. 2020;75:104256.

Sadovy de Mitcheson Y, To AWL, Wong NW, Kwan HY, Bud WS. Emerging from the murk: Threats, challenges and opportunities for the global swim bladder trade. *Rev Fish Biol Fish*. 2019;29:809-35.

Giusti A, Malloggi C, Tinacci L, Nucera D, Armani A. Mislabeling in seafood products sold on the Italian market: A systematic review and meta-analysis. *Food Control*. 2023;145:109395.

Sun Y, Pei Y, Li Q, Xia F, Xu D, Shen G, et al. An expert system for species discrimination and grade identification of fish maw. *Microchem J*. 2024;207:112141.

Kendall H, Naughton P, Kuznesof S, Raley M, Dean M, Clark B, et al. Food fraud and the perceived integrity of European food imports into China. *PLoS ONE*. 2018;13(5):e0195817.

Böhme K, Calo-Mata P, Barros-Velázquez J, Ortea I. Recent applications of omics-based technologies to main topics in food authentication. *TrAC Trends Anal Chem*. 2019;110:221-32.

Zhu SQ, Shi YQ, Huang CH, Cao CF, Ji F, Luo BZ, et al. Research progress of identification techniques for original species of isinglass. *J Food Saf Qual*. 2022;13:3593-601.

Yin H, Yang Q, Huang F, Li H, Wang H, Zheng H, et al. Multimodal fish maw type recognition based on Wasserstein generative adversarial network combined with gradient penalty and spectral fusion. *Spectrochim Acta A Mol Biomol Spectrosc*. 2025;327:125430.

Sha XM, Jiang WL, Hu ZZ, Zhang LJ, Xie ZH, Lu L, et al. Traceability and identification of fish gelatin from seven cyprinid fishes by high performance liquid chromatography and high-resolution mass spectrometry. *Food Chem*. 2023;400:133961.

Silva AJ, Kawalek M, Williams-Hill DM, Hellberg RS. PCR cloning combined with DNA barcoding enables partial identification of fish species in a mixed-species product. *Front Ecol Evol*. 2020;8:28.

Xing B, Chen X, Wu Q, Wang Y, Wang C, Xiang P, et al. Species authentication and conservation challenges in Chinese fish maw market using Mini-DNA barcoding. *Food Control*. 2025;167:110779.

Zareef M, Arslan M, Hassan MM, Ahmad W, Ali S, Li H, et al. Recent advances in assessing qualitative and quantitative aspects of cereals using nondestructive techniques: A review. *Trends Food Sci Technol*. 2021;116:815-28.

Nimbkar S, Auddy M, Manoj I, Shanmugasundaram S. Novel techniques for quality evaluation of fish: A review. *Food Rev Int*. 2023;39:639-62.

Wu K, Ji Z, Wang H, Shao X, Li H, Zhang W, et al. A Comprehensive Review of AI Methods in Agri-Food Engineering: Applications, Challenges, and Future Directions. *Electronics*. 2025;14:3994.

Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2014 Jun 23-28; Columbus, OH, USA. p. 580-7.

Girshick R. Fast R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2015 Dec 7-13; Santiago, Chile. p. 1440-8.

Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:1137-49.

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: Single shot multibox detector. In: *Proceedings of the Computer Vision – ECCV 2016: 14th European Conference; 2016 Oct 11-14; Amsterdam, The Netherlands*. Berlin/Heidelberg: Springer; 2016. p. 21-37.

Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22-29; Venice, Italy*. p. 2980-8.

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; 2016 Jun 27-30; Las Vegas, NV, USA*. p. 779-88.

Varghese R, Sambath M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In: *Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS); 2024 Apr 18-19; Chennai, India*. p. 1-6.

Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: Real-time end-to-end object detection. *Adv Neural Inf Process Syst*. 2024;37:107984-11001.

Khanam R, Hussain M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv*. 2024;arXiv:2410.17725.

Tian Y, Ye Q, Doermann D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv*. 2025;arXiv:2502.12524.

Geng J, Liu J, Kong X, Shen B, Niu Z. The fishmeal adulteration identification based on microscopic image and deep learning. *Comput Electron Agric*. 2022;198:106974.

Kong X, Yu J, Guo J, Tian Q, Li F, Liu Y. AM-YOLO: An Efficient Object Detection Model for *Astragalus mongholicus* Authentication with Dynamic Gated Feature Fusion. *IEEE Access*. 2025;13:163731-48.

Jubayer MF, Ruhad FM, Kayshar MS, Rizve Z, Alam Soeb MJ, Izlal S, et al. Detection and Identification of Honey Pollens by YOLOv7: A Novel Framework toward Honey Authenticity. *ACS Agric Sci Technol*. 2024;4:747-58.

Zhang Y, Xing X, Zhu L, Li X, Wang J, Du Y, et al. A rapid identification technique for rice adulteration based on improved YOLOv8 model. *Meas Sci Technol*. 2025;36:026207.

Jose R, Ponmozhi K. Modified Residual Variable Correlation Kernel Convolutional Neural Network for Classifying Spices. *Anal Lett*. 2025;58:3133-54.

Zoph B, Cubuk ED, Ghiasi G, Lin TY, Shlens J, Le QV. Learning data augmentation strategies for object detection. In: *Proceedings of the Computer Vision – ECCV 2020: 16th European Conference; 2020 Aug 23-28; Glasgow, UK. Part XXVII*. Berlin/Heidelberg: Springer; 2020. p. 566-83.

Wang A, Chen H, Lin Z, Han J, Ding G. Repvit: Revisiting mobile CNN from vit perspective. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16-22; Seattle, WA, USA*. p. 15909-20.

Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning. In: *Proceedings of the ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing; 2023 Jun 4-10; Rhodes Island, Greece*. p. 1-5.

Yu W, Luo M, Zhou P, Si C, Zhou Y, Wang X, et al. Metaformer is actually what you need for vision. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19-23; New Orleans, LA, USA*. p. 10819-29.

Shi D. Transnext: Robust foveal visual perception for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16-22; Seattle, WA, USA. p. 17773-83.

Feng C, Zhong Y, Gao Y, Scott MR, Huang W. Tood: Task-aligned one-stage object detection. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11-17; Montreal, BC, Canada. p. 3490-9.

Zhang H, Zhang S. Shape-IoU: More Accurate Metric considering Bounding Box Shape and Scale. arXiv. 2024;arXiv:2312.17663.

Tong Z, Chen Y, Xu Z, Yu R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. arXiv. 2023;arXiv:2301.10051.

Wang CY, Yeh IH, Mark Liao HY. YOLOv9: Learning what you want to learn using programmable gradient information. In: Proceedings of the European Conference on Computer Vision; 2024 Sep 29-Oct 4; Milan, Italy. p. 1-21.

Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. Detrs beat Yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16-22; Seattle, WA, USA. p. 16965-74.

Dai X, Chen Y, Xiao B, Chen D, Liu M, Yuan L, et al. Dynamic head: Unifying object detection heads with attentions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19-25; Nashville, TN, USA. p. 7373-82.

Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17-24; Vancouver, BC, Canada. p. 7464-75.

Yu Z, Huang H, Chen W, Su Y, Liu Y, Wang X. Yolo-facev2: A scale and occlusion aware face detector. Pattern Recognit. 2024;155:110714.

Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2020 Feb 9-11; New York, NY, USA. Vol. 34. p. 12993-3000.

Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15-20; Long Beach, CA, USA. p. 658-66.

Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing. 2022;506:146-57.

Chen Z, Chen K, Lin W, See J, Yu H, Ke Y, et al. Piou loss: Towards accurate oriented object detection in complex environments. In: Proceedings of the Computer Vision – ECCV 2020: 16th European Conference; 2020 Aug 23-28; Glasgow, UK. Part V. Berlin/Heidelberg: Springer; 2020. p. 195-211.

He Y, Song K, Meng Q, Yan Y. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Trans Instrum Meas. 2019;69:1493-504.