

XraySafe-YOLO: Compact Scale-Aware Fusion and Contour-Motivated Attention for Recall-Oriented Detection of Occluded Prohibited Items in X-ray Images

Liu Liu, Wei Huang*

Department of General Education, Shanghai Customs College, Shanghai, China

Abstract—Occlusion, weak texture, and scale variation make prohibited-item detection difficult in X-ray security images. This study presents XraySafe-YOLO, a compact detector built on YOLO11n that combines Contour-Guided Occlusion Attention with Lightweight Scale-Aware Feature Fusion. On OPIXray, at a fixed confidence threshold of 0.25 and intersection-over-union threshold of 0.50, the full model reduced false negatives from 201 to 163 in the seed-42 experiment. Across three split-and-training trials, the full model reduced false negatives and improved fixed recall over YOLO11n in every trial, with a mean false-negative count of 172.7 and a mean fixed recall of 0.9030. A supplemental single-seed evaluation on the 12-class PIDray benchmark increased mean average precision at an intersection-over-union threshold of 0.50 from 0.7938 to 0.8035 and reduced false negatives from 10,148 to 9,318. LSAFF produced the more stable individual improvement, whereas CGOA alone was sensitive to the OPIXray trial seed. Results on both datasets show the same main behavior: the proposed combination improves recall and reduces missed detections, while some false-positive and precision costs remain. The method is therefore positioned as a compact recall-oriented option for human-assisted X-ray security inspection rather than as a detector that improves every metric.

Keywords—X-ray imaging; prohibited item detection; object detection; YOLO; occlusion-aware detection; multi-scale feature fusion; recall-oriented evaluation

I. INTRODUCTION

X-ray imaging is widely used to inspect baggage, parcels, and cargo at airports, ports, railway stations, logistics centers, and other security checkpoints. It allows inspectors to check concealed objects without opening the container. However, detection becomes difficult when prohibited items are small, have weak contrast, or overlap with normal objects. A high inspection workload can also increase the chance of missed detections.

Automatic detection can help inspectors by marking regions that need further checking. X-ray images are different from normal color images. They have limited texture, incomplete object boundaries, and overlapping materials. Downsampling may also weaken small-object features. These conditions make false negatives an important issue.

Mean average precision is useful for comparing detectors, but it does not show how many prohibited items are missed at a selected threshold. In a human-assisted inspection process, a false positive usually leads to another check, while a false negative may allow a dangerous item to pass. Therefore, this study reports standard detection metrics together with true-positive, false-positive, and false-negative counts, fixed-threshold recall, a recall-weighted F_2 score, occlusion-level results, and model complexity.

YOLO detectors are suitable for inspection tasks because they use a compact one-stage structure. Existing X-ray detection methods often use attention, feature-pyramid changes, larger receptive fields, or efficient convolution. This study focuses on two points. First, local shape and boundary information may still be useful when texture is weak or an object is partly hidden. Second, features from different scales should be combined without making the detector too large.

This study asks three questions. Can compact multi-scale fusion reduce missed detections without replacing the YOLO11n backbone? Does local boundary-motivated attention help by itself, or mainly when it is combined with multi-scale features? Are the results similar across repeated split-and-training trials? We first test the modules separately and together on OPIXray. We then use PIDray to check whether the same behavior appears with more object categories and the official Easy, Hard, and Hidden subsets.

XraySafe-YOLO uses two plug-in modules. Contour-Guided Occlusion Attention (CGOA) applies spatial and channel recalibration to the higher-resolution P3 and P4 features. It is motivated by local boundary changes, but it does not use contour labels or an external edge detector. Lightweight Scale-Aware Feature Fusion (LSAFF) aligns P3, P4, and P5 and learns the contribution of each input scale to each output scale. The modules are tested separately and together.

The main contributions are as follows:

- A compact YOLO11n-based detector is proposed for occluded prohibited items in X-ray images. CGOA applies local spatial reweighting motivated by boundary variation, and LSAFF fuses features for each detection scale.
- A controlled OPIXray ablation compares the two

*Corresponding author.

modules separately and together, with CBAM-lite as an attention reference. The four main configurations are tested in three trials with different validation and training seeds.

- A single-seed second-benchmark test on PIDray compares YOLO11n, YOLO11n + LSAFF, and the full model on the official Easy, Hard, and Hidden subsets.
- The evaluation separates class-averaged standard metrics from micro-averaged fixed-count metrics, so the trade-off between missed detections and false alarms can be seen directly.
- The study reports occlusion-level, class-level, complexity, repeated-trial, operating-point, second-benchmark, and qualitative analyses. Model definitions, configuration files, result summaries, and a small execution demo are available from the corresponding author upon reasonable request.

The remainder of the study is organized as follows. Section II reviews related work. Section III describes the architecture and evaluation design. Section IV presents the experimental protocol, Section V reports the results, and Section VI discusses their practical meaning and limitations. Section VII concludes the study.

II. RELATED WORK

A. Benchmarks and Overlap-Aware X-ray Detection

X-ray prohibited-item detection is different from natural-image detection because overlapping materials, weak texture, and incomplete boundaries can mix object and background information. Public datasets have played an important role in this field. SIXray introduced a large-scale dataset with serious class imbalance and overlapping baggage contents [1]. OPIXray focuses on five cutter-like categories and provides three official occlusion-level test subsets [2], [3]. PIDray includes more categories and hidden objects in realistic scenes [4]–[6].

Recent methods also address overlap directly. AO-DETR uses category-specific assignment and dense decoder refinement to reduce feature coupling and unclear boundaries on PIXray and OPIXray [7]. Dual-view detection shows that two scanner views can help solve cases that remain difficult in one view [8]. These methods are important references, but they use different architectures and, for dual-view detection, different input data from the compact single-view setting used in this study.

B. Lightweight YOLO Detectors for Security X-ray Images

YOLO detectors are widely used when deployment speed and implementation simplicity matter [9]–[13]. X-ray variants have explored attention, feature aggregation, dynamic heads, lightweight backbones, loss redesign, pruning, and distillation. ScanGuard-YOLO emphasizes recall under cluttered and occluded conditions [14]. Fang et al. reduce YOLOv8 channels and introduce a simplified BiFPN to balance accuracy and computation [15]. Fine-YOLO redesigns feature aggregation and uses normalized Wasserstein distance to improve small-object localization [16]. Attention-based X-ray detection has

also combined multi-scale features with explicit feature recalibration [17].

The design space has continued to expand. A lightweight YOLOv7-tiny model targets portable inspection devices through efficient architectural changes [18]. Guan et al. combine deformable convolution and a dynamic detection head in YOLOv8 [19]. Chen et al. report another improved YOLOv8 framework for contraband detection [20]. Xray-YOLO-Mamba integrates selective state-space blocks and evaluates across CLCXray, OPIXray, and SIXray [21]. More recent work combines depthwise convolution, channel-spatial attention, and supervised contrastive learning on two X-ray datasets [22], while GSA-YOLO uses structured sparsity and adaptive knowledge distillation to emphasize both accuracy and real-time efficiency [23].

These studies show that a compact X-ray detector should not be judged only by parameter count. Recent work often reports module ablations, computation, visual examples, and tests on more than one dataset or more than one training run. This study has a narrower goal. It examines how local contour-motivated attention and one-stage scale-weighted fusion affect missed detections, false alarms, and trial-to-trial variation in a compact YOLO11n detector.

C. Efficient Attention and Multi-Scale Fusion

Efficient detectors often use depthwise separable convolution and compact channel transformations derived from MobileNet-style designs [24], [25]. Attention modules such as the Convolutional Block Attention Module (CBAM) and coordinate attention recalibrate spatial or channel responses with limited overhead [26], [27]. In X-ray images, this can help emphasize small visible fragments, although clutter can also create strong responses that do not belong to a prohibited item.

Multi-scale fusion addresses the complementary need for fine spatial detail and broader context. Feature Pyramid Networks (FPNs) and Path Aggregation Networks (PANets) exchange information between feature levels, EfficientDet uses repeated bidirectional weighted fusion, and adaptively spatial feature fusion (ASFF) learns scale contributions at the spatial level [28]–[31]. LSAFF follows the same broad motivation but uses one normalized weighted fusion for each output scale rather than a repeated bidirectional pyramid. CGOA is likewise deliberately compact and requires neither contour annotations nor an external edge detector.

D. Evaluation Gap and Positioning of this Study

Most X-ray detection papers report mAP, precision, recall, parameters, FLOPs, and selected examples. These measures are useful, but they may not show how many prohibited items are missed at one selected threshold or whether an improvement is stable across training runs. In a human-assisted workflow, it is also important to know whether lower FN is obtained with more FP.

For this reason, the study reports standard metrics together with class-aware fixed-threshold TP, FP, and FN counts, a recall-weighted F_2 score, occlusion-level and class-level results, three OPIXray trials for the four main configurations,

and a PIDray second-benchmark test. Table I summarizes representative recent work. The absolute scores are not compared across rows because the datasets, classes, resolutions, and evaluation settings are different.

III. METHODOLOGY

A. Overview of XraySafe-YOLO

XraySafe-YOLO adds two low-overhead modules to YOLO11n. The modules address incomplete local structure and large changes in object scale. The network keeps the standard one-stage detection pipeline and the original detection heads.

An input image is resized and normalized before entering the backbone. The backbone and neck produce three feature levels, P3, P4, and P5. CGOA is applied to P3 and P4, where spatial detail is still sufficient to represent small or partly visible object boundaries; P5 is left unchanged. In the CGOA-only model, the enhanced P3 and P4 features proceed to the detection heads through the original neck. In the combined model, P3', P4', and P5 are passed to LSAFF, which returns $P3^{out}$, $P4^{out}$, and $P5^{out}$ for the three heads. The heads predict bounding boxes, categories, and confidence scores.

Both modules are designed as plug-ins rather than as a replacement for the YOLO11n backbone. Fig. 1 summarizes the data flow.

B. Baseline Detector and Model Variants

YOLO11n is used as the compact baseline. All variants retain its backbone, neck, and detection heads, which keeps the comparison focused on the added modules. Five configurations are evaluated at seed 42: YOLO11n, YOLO11n + CBAM-lite, YOLO11n + CGOA, YOLO11n + LSAFF, and YOLO11n + CGOA + LSAFF. The four principal configurations excluding CBAM-lite are also evaluated with seeds 42, 2026, and 3407.

The implementation places CGOA after the final P3 and P4 neck branches; P5 bypasses the module. In the model definition, these features correspond to layers 16, 19, and 22. LSAFF is instantiated once for each target resolution so that the same P3/P4/P5 inputs produce one fused output for each detection head. Shape-compatible YOLO11n parameters are transferred from the pretrained checkpoint to each custom configuration, while parameters introduced by CGOA, LSAFF, and CBAM-lite use their module initializations. Input size, batch size, epoch limit, data preparation, optimizer policy, augmentation, early stopping, and fixed-count evaluation settings are kept the same across models; no architecture-specific hyperparameter tuning is performed.

C. Contour-Guided Occlusion Attention Module

X-ray images usually contain less texture than natural images. However, the shape and boundary of a prohibited item may still be visible when the item is partly covered. For example, part of a knife blade or the outline of scissors can still provide useful information in cluttered baggage. CGOA is based on this observation. In the module name, “guided” means that the design is guided by local boundary variation; it does not use an external contour image or contour supervision.

Let $F \in \mathbb{R}^{C \times H \times W}$ denote an intermediate feature map extracted by the backbone. CGOA is inserted into the P3 and P4 branches because these feature levels preserve more spatial detail and are more relevant to small or partially occluded prohibited items. The module first generates a local spatial response map from F . This response is contour-motivated rather than contour-supervised: no edge labels, contour annotations, or external edge detector are used. Specifically, the spatial branch applies a 3×3 depthwise convolution with batch normalization and SiLU activation, followed by a 1×1 convolution that reduces the response to one channel. The contour branch produces a spatial logit map $M_c \in \mathbb{R}^{1 \times H \times W}$.

The contour response is then combined with channel-wise recalibration. Given the feature map F , global average pooling and global max pooling are used to obtain channel descriptors. The two descriptors pass through the same two-layer 1×1 convolutional MLP with SiLU activation. Its hidden width is $\max(\lfloor C/16 \rfloor, 8)$. The summed MLP output is denoted by L_{ch} , and the channel weights are $A_{ch} = \sigma(L_{ch}) \in [0, 1]^{C \times 1 \times 1}$.

The enhanced feature map is computed as follows:

$$F' = F \otimes (1 + \sigma(M_c)) \otimes A_{ch}, \quad (1)$$

where, $\otimes$ denotes element-wise multiplication with broadcasting and $\sigma(\cdot)$ denotes the sigmoid activation function. The residual spatial factor $1 + \sigma(M_c)$ is never smaller than one, so the spatial branch alone does not attenuate F ; however, the subsequent channel gate $A_{ch} \in [0, 1]$ can attenuate channels in the final product. CGOA is restricted to P3 and P4 because these feature maps retain finer boundary information; applying the same spatial gate to the coarser P5 branch would add computation where local contour detail is already strongly downsampled.

The residual spatial gate is intended to emphasize local transitions without directly suppressing them; the complete CGOA output can still be attenuated by the channel gate. Thus, boundary enhancement is treated as a design motivation rather than as a demonstrated contour detector. Because X-ray clutter can also create strong local transitions, the ablation study evaluates whether this bias helps consistently rather than assuming that it will do so. Fig. 2 illustrates the CGOA module.

D. Lightweight Scale-Aware Feature Fusion Module

Prohibited items can have different sizes and irregular shapes. Small items need high-resolution details, while large or heavily occluded items need more context. Multi-scale feature fusion is therefore useful. However, some fusion structures add many parameters and computations, which is not suitable for a compact baseline.

LSAFF addresses this trade-off with one output-specific weighted fusion stage for each detection scale. For every input-to-output path, the implementation uses the same architectural pattern: a 3×3 depthwise convolution followed by a 1×1 pointwise convolution, with batch normalization and SiLU activation after each convolution. The resulting feature is resized to the target spatial resolution with nearest-neighbor interpolation. The three output-specific LSAFF instances have

TABLE I. REPRESENTATIVE RECENT DIRECTIONS IN X-RAY PROHIBITED-ITEM DETECTION AND THE POSITION OF THIS STUDY

Method	Year	Evaluation Data	Main Technical Emphasis	Evidence Emphasized
AO-DETR [7]	2025	PIXray, OPIXray	Category-specific assignment and dense decoder refinement for overlap	Multi-dataset accuracy, overlap analysis, qualitative localization
Lightweight YOLOv8 [15]	2024	PIDray, CLCXray	Channel reduction, simplified BiFPN, and lightweight loss design	Accuracy-complexity balance and second-benchmark validation
Fine-YOLO [16]	2024	EDS, NEU-DET	Lightweight feature aggregation and normalized Wasserstein distance	Small-object accuracy, complexity, auxiliary validation
Xray-YOLO-Mamba [21]	2025	CLCXray, OPIXray, SIXray	State-space feature extraction and lightweight sampling blocks	Three-dataset results, FLOPs, parameters, FPS, heatmaps
YOLOv8-DWConv-CSAF [22]	2026	CLCXray, HiXray	Depthwise convolution, channel-spatial attention, contrastive learning	Two-dataset validation, efficiency, multi-scale analysis
GSA-YOLO [23]	2026	HiXray, PIDray	Structured sparsity and adaptive knowledge distillation	Cross-dataset accuracy and real-time inference
XraySafe-YOLO	–	OPIXray, PIDray	CGOA, LSAFF, and controlled YOLO1n ablations	Fixed TP/FP/FN, three-trial variability, occlusion, class, and second-benchmark analysis

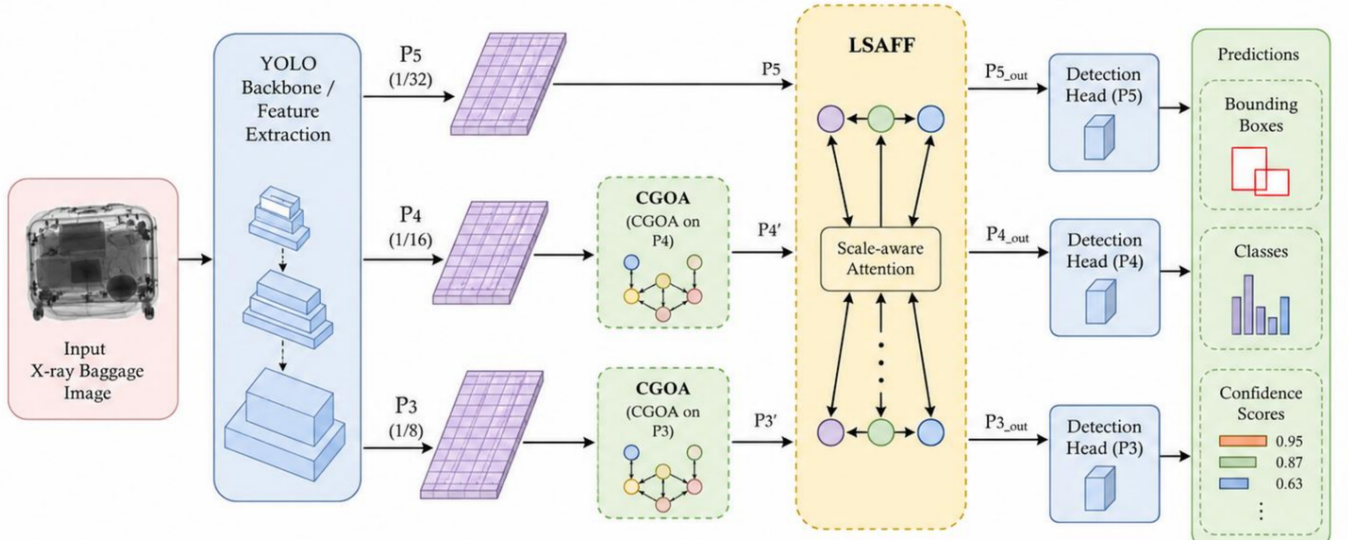


Fig. 1. Overall architecture of the proposed XraySafe-YOLO framework. The YOLO backbone extracts three feature maps, P3, P4, and P5. CGOA is applied only to the P3 and P4 branches to reweight local spatial responses, motivated by boundary transitions in small and partially occluded prohibited items, while P5 bypasses CGOA and enters LSAFF directly. LSAFF then produces learned scale-aware output features $P3^{out}$, $P4^{out}$, and $P5^{out}$ for the three detection heads, which output bounding boxes, classes, and confidence scores.

independent parameters; the transformations are not shared across target scales.

Let F_i denote the feature map from the i -th input scale, where $i \in \{3, 4, 5\}$. Unlike a single-output fusion block, LSAFF produces three scale-specific output features for the detection heads. The transformation from input scale i to output scale k is denoted by $\phi_{k,i}(\cdot)$. For each output scale $k \in \{3, 4, 5\}$, LSAFF assigns normalized learned scale weights $\alpha_{k,i}$ to the aligned input features. Let the weighted intermediate fusion be:

$$\tilde{F}_k = \sum_{i \in \{3, 4, 5\}} \alpha_{k,i} \cdot \phi_{k,i}(F_i), \quad k \in \{3, 4, 5\}, \quad (2)$$

and let $\psi_k(\cdot)$ denote the output-specific depthwise-separable refinement block. The feature delivered to the k -th detection head is

$$F_k^{out} = \psi_k(\tilde{F}_k). \quad (3)$$

Here, $\alpha_{k,i}$ is a learned scalar contribution of the i -th input scale to the k -th output scale. Each output-specific LSAFF instance owns an independent three-element logit vector initialized to zero, corresponding to equal initial contributions from P3, P4, and P5. The weights are normalized for each output scale:

$$\alpha_{k,i} = \frac{\exp(w_{k,i})}{\sum_{j \in \{3, 4, 5\}} \exp(w_{k,j})}. \quad (4)$$

The refinement ψ_k is a second 3×3 depthwise-separable convolution applied after weighted summation. The weights are learned independently for each output scale; they are scalar coefficients rather than image-dependent or spatially

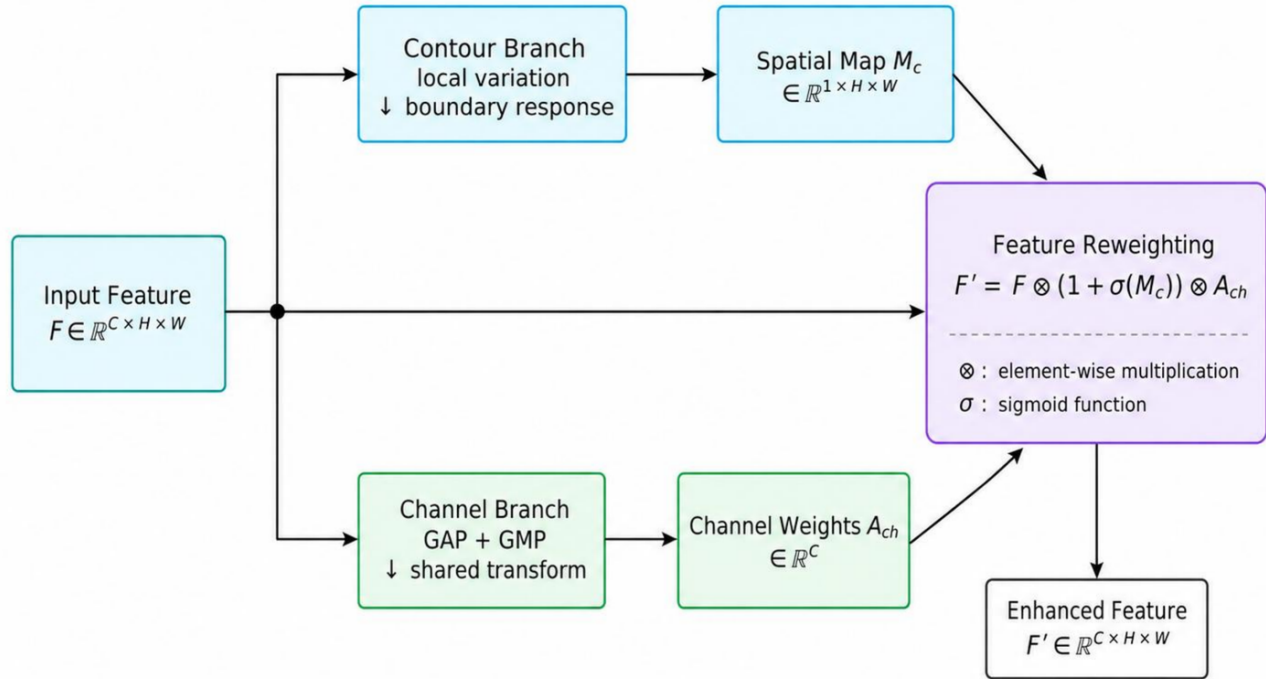


Fig. 2. Contour-Guided Occlusion Attention (CGOA) module. Given an input feature map F , the contour-motivated spatial branch generates a response map M_c , the channel branch generates channel weights A_{ch} using global average and max pooling, and the identity path preserves the original feature. The enhanced output is computed as $F' = F \otimes (1 + \sigma(M_c)) \otimes A_{ch}$ without requiring additional contour annotations.

varying attention maps. High-resolution features retain small-object and boundary information, while deeper features provide broader context. Unlike bidirectional pyramids such as BiFPN, which repeatedly propagate features in top-down and bottom-up paths, LSAFF performs one weighted fusion for each output scale. This keeps the added computation modest, although no direct LSAFF–BiFPN performance comparison is made in this study. Fig. 3 illustrates the module.

E. Training Objective and Recall-Oriented Evaluation

The model uses the standard YOLO detection loss. This keeps the comparison focused on the proposed modules and avoids adding another change through the loss function. The training loss follows the common YOLO-style structure:

$$\mathcal{L} = \lambda_{box} \mathcal{L}_{box} + \lambda_{cls} \mathcal{L}_{cls} + \lambda_{dfl} \mathcal{L}_{dfl}, \quad (5)$$

where, $\mathcal{L}_{box}$ is the bounding-box regression loss, $\mathcal{L}_{cls}$ is the classification loss, and $\mathcal{L}_{dfl}$ is the distribution focal loss used by the underlying YOLO training framework [32]. No class-balanced or hard-sample loss is added. This design keeps the ablation focused on the two proposed architectural modules, CGOA and LSAFF.

Because prohibited item detection is sensitive to missed detections, recall orientation is reflected in the evaluation protocol rather than in a separate training loss. In addition to mAP, the analysis reports recall, false-negative count, fixed-threshold recall, occlusion-level recall, and efficiency metrics. These metrics make it possible to determine whether the proposed modules improve difficult X-ray inspection cases

without overstating the model as an autonomous safety decision system.

IV. EXPERIMENTAL PROTOCOL

A. Datasets

The primary experiments use the public OPIXray benchmark, which was constructed for occluded prohibited item detection in X-ray security inspection images [2], [3]. OPIXray contains 8885 images from five cutter-like prohibited item categories: folding knife, straight knife, scissors (official label: “scissor”), utility knife, and multi-tool knife. The official test set is divided into three occlusion-level subsets, denoted as OL1, OL2, and OL3, which support a direct analysis of detection behavior under increasing overlap and visual clutter. Table II summarizes the OPIXray split used for the controlled ablation and repeated trials.

A supplemental evaluation uses the current 124,486-image PIDray release distributed through the official lutao2021/PIDray repository [4]–[6]. This release contains 76,913 training images and official Easy, Hard, and Hidden test subsets with 24,758, 9,746, and 13,069 images, respectively, across 12 prohibited-item categories. A deterministic 5% holdout from the official training split is used for validation, while all images in the three official test subsets are retained, including images without prohibited-item boxes; predictions on such images therefore contribute false positives when applicable. The converted test annotations contain 8450, 14,500, and 19,050 ground-truth boxes, respectively, for a total of 42,000. Because the models are trained on PIDray training data, this experiment evaluates the architecture on a

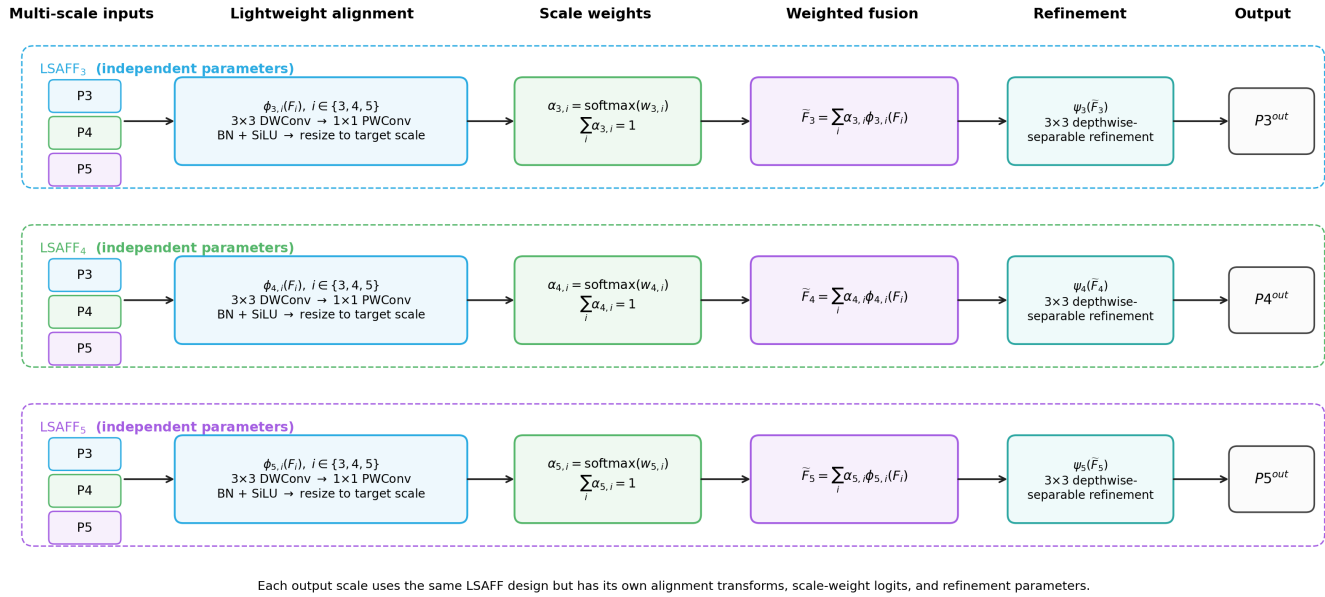


Fig. 3. Lightweight Scale-Aware Feature Fusion (LSAFF) module. Multi-scale inputs P3, P4, and P5 are aligned and transformed through lightweight operations. Three independently parameterized output-specific blocks align P3/P4/P5, learn normalized scale weights $\alpha_{k,i}$, form $\bar{F}_k$, and apply the refinement ψ_k before producing $P3^{out}$, $P4^{out}$, and $P5^{out}$.

second benchmark rather than cross-dataset generalization of OPIXray-trained weights.

Representative OPIXray examples are shown in Fig. 4. The examples illustrate why prohibited item detection in X-ray images is more difficult than ordinary natural-image detection: item contours can be weak, objects can be transparent or overlapping, and cluttered baggage contents may visually interfere with cutter-like objects.

B. Data Preparation

OPIXray is converted into a unified YOLO-compatible annotation format. Each image is associated with a text annotation file containing the object category and normalized bounding-box coordinates. The official OPIXray test set is preserved in every trial. For the seed-42 architecture comparison, 711 images are selected from the official training portion using split seed 42. For the repeated seeded evaluation, the 711-image validation subset is regenerated using split seeds 42, 2026, and 3407, respectively; the official test set remains unchanged.

The conversion process includes parsing original annotation files, mapping dataset-specific categories to unified class names, converting bounding-box coordinates to normalized YOLO format, verifying image-annotation consistency, generating train/validation/test file lists, and checking invalid boxes, missing images, and duplicated annotations.

PIDray is converted from its COCO-style detection annotations to the same YOLO-compatible format while retaining all 12 official categories. The training images are partitioned deterministically into training and validation sets with a 5% validation ratio and seed 42. Easy, Hard, and Hidden remain separate test sets. The combined PIDray standard metrics are obtained by running the validator once on the union of all three

official test subsets (pooled predictions), not by averaging the three subset mAP values. Fixed TP/FP/FN counts are summed over the same pooled test population.

C. Implementation Details

The experiments were run on Windows with Python 3.11, PyTorch, Ultralytics YOLO 8.4.78, and an NVIDIA GeForce RTX 4070 SUPER GPU with 12 GB graphics memory. Images were resized to 640×640 , the batch size was 48, and training was allowed to continue for up to 100 epochs with early-stopping patience 30. All models used pretrained YOLO11n initialization, cosine learning-rate scheduling, weight decay 0.0005, three warm-up epochs, deterministic seed settings, and mixed-precision training. Standard Ultralytics augmentations were retained, including HSV adjustment, translation, scaling, horizontal flipping, and mosaic augmentation with `close_mosaic=10`. No separate RandAugment or random-erasing pipeline was added.

The same Ultralytics training interface and `optimizer=auto` policy were used for every model, and no model received architecture-specific optimizer tuning. Under Ultralytics 8.4.78, the OPIXray runs resolved to AdamW with initial learning rate 0.001111 and $\beta_1 = 0.9$; cosine scheduling used the framework default final learning-rate factor 0.01. Weight decay was 0.0005 and warm-up lasted three epochs. The larger PIDray runs resolved to MuSGD with initial learning rate 0.01 and momentum 0.9 under the same automatic-selection rule. Shape-compatible pretrained parameters and newly initialized module parameters were optimized together. These resolved settings are reported to make clear that `optimizer=auto` was not an unspecified degree of freedom.

The seed-42 OPIXray comparison used split seed 42 and training seed 42 for all five models. For the repeated OPIXray

TABLE II. OPIXRAY DATASET STATISTICS USED FOR THE PRIMARY EXPERIMENTS

Dataset	Images	Classes	Train	Val	Test	Test GT Boxes by OL Subset
OPIXray	8885	5	6398	711	1776	OL1: 923; OL2: 550; OL3: 307



Fig. 4. Representative OPIXray examples from OL1, OL2, and OL3 occlusion-level subsets. The images show the transparent appearance, weak texture, and object overlap that motivate contour-motivated and recall-oriented detection analysis.

evaluation, the four principal models were trained with seeds 42, 2026, and 3407. In each trial, the validation split seed matched the training seed, while the official OPIXray test set remained unchanged. The repeated results therefore combine split and optimization variability; they are not a fixed-split significance test. Hardware-dependent validation timing was recorded only as a diagnostic and is not used to support the architectural conclusions.

The supplemental PIDray experiment used the same software and hardware environment, 640-pixel input resolution, pretrained initialization, augmentation family, optimizer policy, early-stopping mechanism, and fixed-count thresholds. Because of the larger dataset, training used batch size 16 for up to 50 epochs with split and training seed 42. YOLO11n, YOLO11n + LSAFF, and the full model were trained under the same PIDray protocol. The experiment uses one seed and is interpreted as second-benchmark evidence rather than a significance test.

For the seed-42 OPIXray architecture comparison, training completed 100 epochs for YOLO11n and CGOA, 94 epochs for LSAFF, and 92 epochs for the full model because early stopping was triggered for the latter two runs. The reported test metrics use the saved `best.pt` checkpoint selected by the Ultralytics validation fitness, which weights mAP@0.5 and

TABLE III. RESOLVED TRAINING SETTINGS FOR THE PRINCIPAL CONFIGURATIONS

Setting	YOLO11n	CGOA	LSAFF	Full
Optimizer (OPIXray)	AdamW	AdamW	AdamW	AdamW
Initial LR	0.001111	0.001111	0.001111	0.001111
β_1 / momentum	0.9	0.9	0.9	0.9
Weight decay	0.0005	0.0005	0.0005	0.0005
Completed epochs (seed 42)	100	100	94	92
Checkpoint	best validation fitness: 0.1 mAP@0.5 + 0.9 mAP@0.5:0.95			

mAP@0.5:0.95 by 0.1 and 0.9, respectively. During fixed-count inference, the prediction confidence threshold is varied as stated in each analysis; non-maximum suppression uses IoU 0.50, class-aware suppression (`agnostic_nms=False`), and `max_det=300`. The subsequent TP/FP/FN matcher separately uses same-class IoU ≥ 0.50 . Thus, the NMS IoU threshold and the matching IoU threshold are numerically equal in this study but serve different operations (Table III).

For PIDray, the same `optimizer=auto` rule resolves to MuSGD because of the larger number of planned iterations; its initial learning rate is 0.01 with momentum 0.9. The PIDray

runs use the same weight decay, cosine scheduling, warm-up policy, checkpoint rule, and inference settings unless otherwise stated.

D. Evaluation Metrics

Standard object detection metrics are reported:

$$Precision = \frac{TP}{TP + FP}, \quad (6)$$

$$Recall = \frac{TP}{TP + FN}, \quad (7)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \quad (8)$$

In addition, mAP@0.5 and mAP@0.5:0.95 are reported. The standard precision, recall, and F1 values are class-averaged metrics returned by the Ultralytics validator. They are therefore not numerically interchangeable with the fixed-count metrics below, which are micro-averaged over all 1780 test instances at one selected confidence and IoU operating point. Per-class recall and AP values are retained in the metric JSON files and summarized for the baseline and full model to inspect whether the aggregate gain is concentrated in only one category.

Because missed prohibited items are important in security inspection, additional recall-oriented metrics are also reported. These include false-negative count, occlusion-level recall, and fixed-threshold recall. Deployment-oriented complexity metrics include parameter count, GFLOPs, and model size. These metrics describe architectural overhead while avoiding unsupported claims about hardware-dependent latency or throughput.

The fixed-count protocol is computed after the model's standard prediction and non-maximum suppression stage. Predictions are filtered with confidence ≥ 0.25 and sorted by descending confidence. Each prediction is greedily matched to the unmatched ground-truth box of the same class with the highest IoU, provided that $\text{IoU} \geq 0.50$. Each ground-truth instance can be matched at most once. Unmatched predictions are counted as false positives, and unmatched ground-truth instances are counted as false negatives. Fixed-count precision and recall are then computed as:

$$Fixed\ Precision = \frac{TP_{fixed}}{TP_{fixed} + FP_{fixed}}, \quad (9)$$

$$Fixed\ Recall = \frac{TP_{fixed}}{TP_{fixed} + FN_{fixed}}. \quad (10)$$

Because missed detections are emphasized in the intended human-in-the-loop screening setting, a recall-weighted fixed-count F_2 score is also reported; with $\beta = 2$, recall receives four times the weight of precision in the denominator-weighted harmonic formulation:

$$F_2 = \frac{5 \cdot Fixed\ Precision \cdot Fixed\ Recall}{4 \cdot Fixed\ Precision + Fixed\ Recall}. \quad (11)$$

For the seed-42 OPIXray test set, this protocol counts 1776 test images and 1780 ground-truth boxes. For PIDray, the same protocol is applied to the 42,000 ground-truth boxes across the official Easy, Hard, and Hidden subsets. The 0.25 confidence threshold is a pre-specified diagnostic point applied identically to every model rather than a claim of universal optimality. Operational deployment would require threshold calibration on data from the target scanner and analysis of the full precision-recall curve.

E. Baseline and Comparative Models

The main OPIXray ablation compares YOLO11n, YOLO11n + CGOA, YOLO11n + LSAFF, and YOLO11n + CGOA + LSAFF. This arrangement isolates the two modules within the same detector and training protocol. CBAM-lite is included at seed 42 as a generic channel-and-spatial attention reference trained under the same settings. The compact PIDray comparison includes YOLO11n, YOLO11n + LSAFF, and the full model. CGOA alone is not repeated on PIDray because the OPIXray trials already identify LSAFF as the stronger standalone component; the second-benchmark experiment focuses on whether the same architectural pattern is observed on a broader 12-class benchmark.

F. Visualization and Descriptive Analysis

The numerical tables are supported by figures generated directly from the reported test results. Error bars in the repeated-trial figures denote the sample standard deviation over seeds 42, 2026, and 3407. The OPIXray occlusion figure uses the fixed FN counts from the official OL1–OL3 subsets, and the class figure plots the recall difference between the full model and YOLO11n. The PIDray second-benchmark figure compares mAP@0.5 and fixed recall across Easy, Hard, and Hidden. These visual summaries do not introduce additional samples or alternative metric definitions; they are intended to make variability, trade-offs, category-specific behavior, and second-benchmark behavior easier to inspect.

G. Reproducibility and Consistency Checks

Dataset split seeds, training seeds, model definitions, conversion summaries, and result summaries were retained for each experiment. A compact code package available from the corresponding author contains the CGOA, LSAFF, and CBAM-lite modules, the associated YAML files, and a CPU-compatible tensor-shape smoke test. The original OPIXray and PIDray images and annotations are not redistributed and must be obtained from their official sources under the applicable academic-use terms.

The fixed-count tables were checked directly from their TP, FP, FN, and ground-truth totals. Every row satisfies $TP + FN = GT$, and fixed precision and recall were recomputed from the counts. For the occlusion analysis, the OL1–OL3 TP and FN values sum to the full 1780-instance test set for each model. These checks place the aggregate and subset TP/FN results on the same ground-truth basis.

V. RESULTS AND ANALYSIS

A. Controlled Seed-42 Architecture Comparison

Table IV reports the conventional seed-42 metrics. The ranking depends on the metric: CBAM-lite has the highest precision, while LSAFF has the highest recall, F1-score, mAP@0.5, and mAP@0.5:0.95. The combined model is close to LSAFF but does not lead any conventional metric. Its class-averaged recall (0.8439) is the arithmetic mean of the five class recalls in Table VII; F1 is recomputed from the corresponding precision and recall. CGOA alone performs below YOLO11n in this trial. Among the proposed individual modules, LSAFF provides the stronger seed-42 result.

The fixed operating-point counts in Table V show the trade-off between missed detections and false alarms more clearly than mAP alone. Relative to YOLO11n, the seed-42 full model detects 38 additional ground-truth instances (1617 versus 1579), reducing FN from 201 to 163, or by 18.9%. This recall gain is accompanied by an increase in FP from 224 to 250, or by 11.6%. At this operating point, each added false positive is associated with about 1.46 additional true detections. Fixed recall rises from 0.8871 to 0.9084, fixed precision decreases from 0.8758 to 0.8661, and the recall-weighted F_2 increases from 0.8848 to 0.8996. The ratio is not a deployment utility measure because the costs of misses and false alarms are application-dependent; it simply makes the observed count trade-off easier to interpret. Therefore, the combined model is better described as a recall-oriented option, not as a model that improves every error type or metric.

At seed 42, adding CGOA to LSAFF reduces FN from 184 to 163 and increases fixed recall from 0.8966 to 0.9084, but FP changes from 255 to 250 and fixed precision remains below the baseline. CGOA alone is worse than YOLO11n in FN and conventional mAP for this seed. The result suggests that CGOA can complement LSAFF at this operating point, but it does not support CGOA as an independently superior attention block.

B. Occlusion-Level and Per-Class Behavior

Table VI reports the seed-42 results on OL1, OL2, and OL3. The three subsets contain 923, 550, and 307 ground-truth instances, respectively, and their TP and FN counts sum exactly to the 1780-instance full test set for every model. The full model gives the lowest FN on OL1 and OL3, whereas LSAFF gives the lowest FN on OL2. The results suggest that multi-scale fusion is the more stable contributor and that the combined configuration is most useful when missed detections receive greater weight.

Fig. 5 visualizes the fixed FN counts. Because OL1, OL2, and OL3 contain different numbers of ground-truth boxes, bar heights should be compared between models within a subset rather than across subsets. Relative to YOLO11n, the full model reduces FN by 20 on OL1, 7 on OL2, and 11 on OL3. In fixed-recall terms, these changes correspond to gains of 0.0217, 0.0127, and 0.0358, respectively. LSAFF records the lowest FN on OL2, while the full model is lowest on OL1 and OL3. The largest proportional recall gain occurs on OL3, although this subset is also the smallest and should not be treated as an independent generalization test.

Table VII reports a compact class-level recall comparison between YOLO11n and the full configuration. The full model improves recall for multi-tool knife, the official “scissor” label, straight knife, and utility knife, but decreases recall for folding knife. The gain is especially visible for multi-tool knife and utility knife, whereas straight knife remains the most difficult category. The table focuses on recall because the class-level analysis concerns missed detections. Supporting class-level outputs are available from the corresponding author upon reasonable request.

Fig. 6 shows that the improvement is different across classes. Multi-tool knife has the largest recall increase (+0.0408), followed by utility knife (+0.0196) and straight knife (+0.0165). The small gain for scissors (+0.0039) occurs on a class that already has high baseline recall. Folding knife decreases by 0.0118, showing that the aggregate gain is not uniform; its compact folded geometry may be confused with dense metallic clutter, although the present results do not isolate a single causal factor. Straight knife remains the most difficult class even after improvement, with recall increasing only from 0.5660 to 0.5825. Its elongated, low-contrast appearance and frequent overlap provide a plausible explanation for the persistent difficulty, but this interpretation should be treated as qualitative rather than causal.

C. Repeated Split-and-Training Evaluation

Table VIII summarizes the repeated split-and-training evidence for all four principal models, and Table IX reports the individual trial values. Each trial uses the same batch size, image size, epoch limit, and evaluation protocol, while the validation split seed equals the training seed. The results are therefore descriptive repeated-holdout evidence rather than a fixed-split significance test.

Table IX reports each trial together with the paired change relative to YOLO11n under the same split-and-training seed. This pairing is descriptive rather than a fixed-split significance test, but it shows directly that LSAFF and the full model reduce FN and raise fixed recall in all three trials.

The three trials show different tendencies for the two modules. CGOA has the largest variation in mAP and FN and does not improve the mean values relative to YOLO11n. LSAFF improves mAP@0.5 and fixed recall and lowers FN in each of the three trials, making it the more consistent standalone module. The combined model also lowers FN and raises fixed recall relative to YOLO11n in every trial, while fixed precision is slightly lower in every trial. Averaged over the three trials, its FN count is 15.9% below the baseline mean (172.7 versus 205.3), fixed recall is higher by 0.0184, and fixed precision is lower by 0.0054. The gain is therefore repeatable in direction but concentrated on recall rather than on all operating-point metrics.

Fig. 7 shows the same pattern in count form. The mean FN decreases by 19.3 for LSAFF and by 32.6 for the full model relative to YOLO11n, whereas CGOA changes the mean by only +0.7. CGOA also has the largest FN standard deviation (31.8), consistent with the poor seed-42 result and stronger results in the other two trials. The error bars overlap, so the figure is descriptive evidence rather than a claim of statistical significance.

TABLE IV. SEED-42 BATCH-48 CONVENTIONAL DETECTION METRICS ON THE OPIXRAY TEST SET. BOLD VALUES INDICATE THE BEST RESULT IN EACH COLUMN

Model	Precision	Recall	F1	mAP@0.5	mAP@0.5:0.95
YOLO11n	0.8744	0.8301	0.8517	0.8852	0.4123
YOLO11n + CBAM-lite	0.8945	0.8372	0.8649	0.8904	0.4146
YOLO11n + CGOA	0.8512	0.7977	0.8232	0.8467	0.3818
YOLO11n + LSAFF	0.8725	0.8588	0.8650	0.8927	0.4151
YOLO11n + CGOA + LSAFF	0.8766	0.8439	0.8599	0.8873	0.4140

TABLE V. SEED-42 FIXED OPERATING-POINT RESULTS AT CONFIDENCE 0.25 AND IOU 0.50. TP, FP, AND FN ARE OBTAINED BY CLASS-AWARE GREEDY MATCHING AFTER NON-MAXIMUM SUPPRESSION

Model	TP	FP	FN	Fixed P	Fixed R	F_2
YOLO11n	1579	224	201	0.8758	0.8871	0.8848
YOLO11n + CBAM-lite	1572	227	208	0.8738	0.8831	0.8812
YOLO11n + CGOA	1540	238	240	0.8661	0.8652	0.8654
YOLO11n + LSAFF	1596	255	184	0.8622	0.8966	0.8895
YOLO11n + CGOA + LSAFF	1617	250	163	0.8661	0.9084	0.8996

TABLE VI. OCCLUSION-LEVEL ANALYSIS ON THE OFFICIAL OPIXRAY TEST SUBSETS. RECALL IS THE STANDARD DETECTOR RECALL COMPUTED ON EACH SUBSET; FN IS COUNTED USING THE FIXED CONFIDENCE/IOU PROTOCOL

Model	OL1 Recall / FN	OL2 Recall / FN	OL3 Recall / FN	OL1 mAP@0.5	OL2 mAP@0.5	OL3 mAP@0.5
YOLO11n	0.8574 / 86	0.8081 / 67	0.7787 / 48	0.9052	0.8726	0.8609
YOLO11n + CGOA	0.8132 / 104	0.7927 / 75	0.7501 / 61	0.8656	0.8408	0.8084
YOLO11n + LSAFF	0.8634 / 87	0.8404 / 58	0.8398 / 39	0.8898	0.8957	0.8840
YOLO11n + CGOA + LSAFF	0.8734 / 66	0.8342 / 60	0.8150 / 37	0.9027	0.8737	0.8774

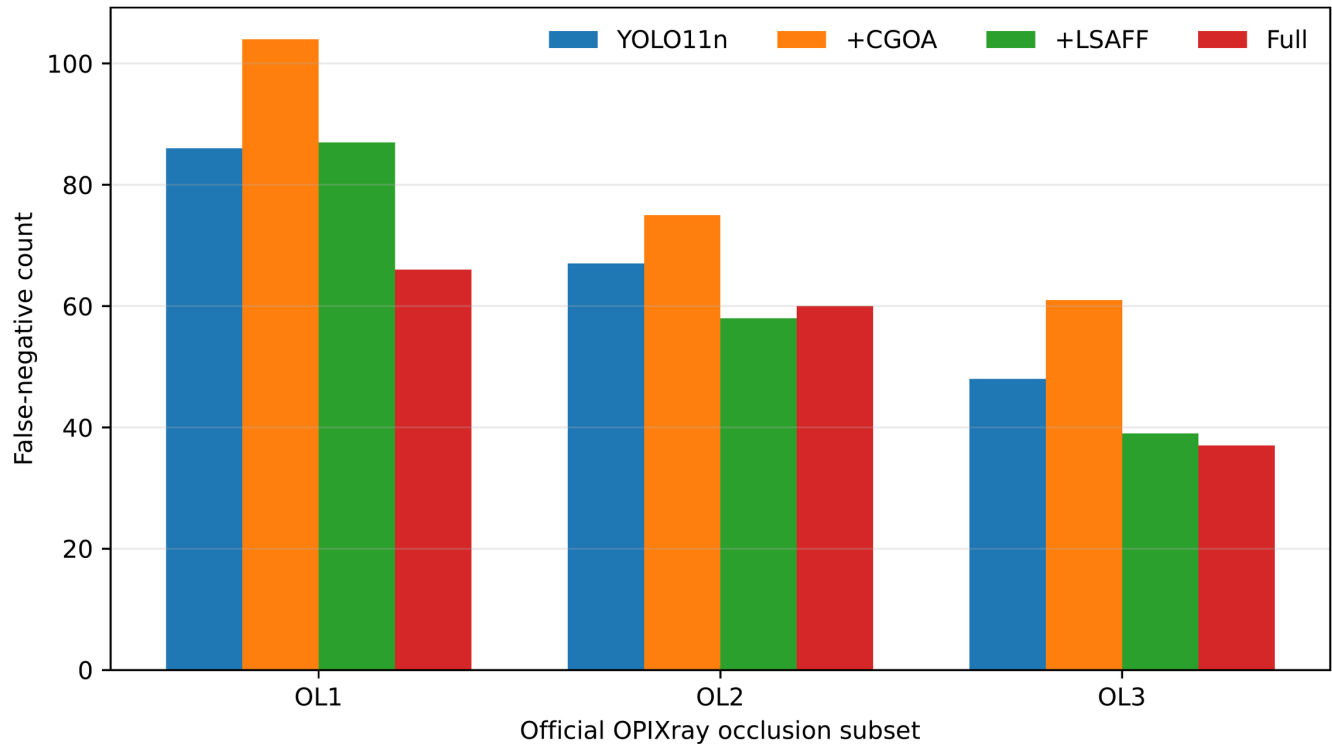


Fig. 5. Fixed false-negative counts on the official OPIXray occlusion subsets at seed 42. The subsets contain different numbers of ground-truth boxes, so comparisons are made between models within each subset.

TABLE VII. CLASS-LEVEL RECALL COMPARISON BETWEEN YOLO11n AND YOLO11n + CGOA + LSAFF ON OPIXRAY

Class	YOLO11n Recall	Full Recall	Δ Recall
folding knife	0.8985	0.8867	-0.0118
multi-tool knife	0.8795	0.9203	0.0408
scissor	0.9755	0.9794	0.0039
straight knife	0.5660	0.5825	0.0165
utility knife	0.8309	0.8505	0.0196

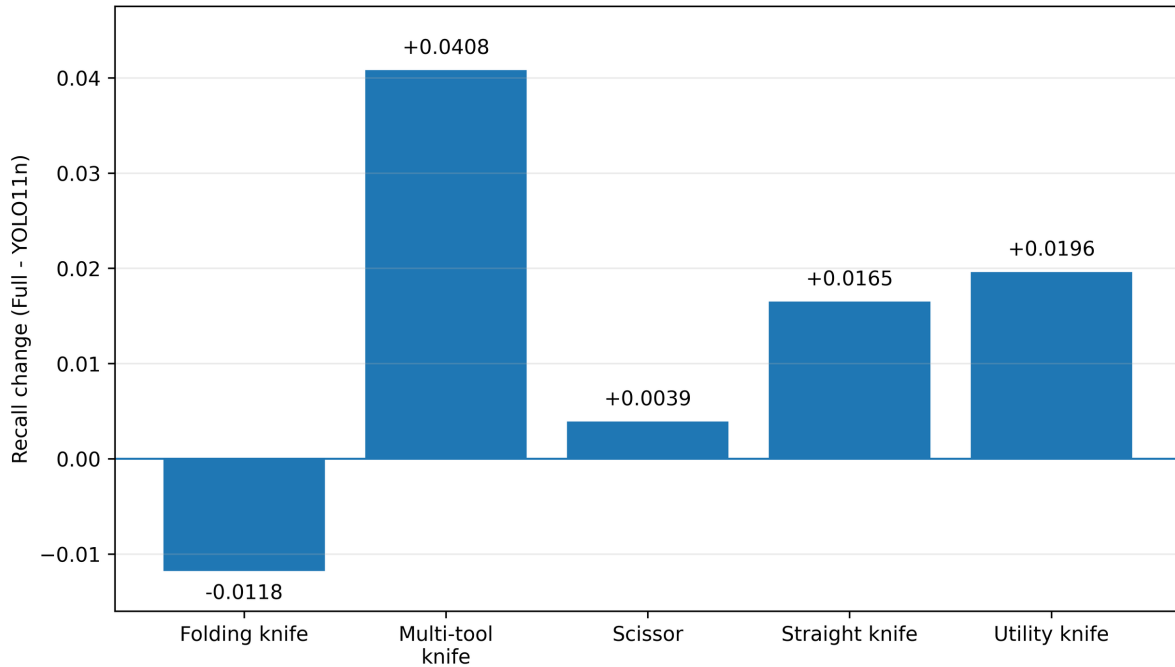


Fig. 6. Class-level recall change of the full model relative to YOLO11n at seed 42. Positive values indicate higher recall for the full model; the folding-knife class shows a small decrease.

TABLE VIII. REPEATED SPLIT-AND-TRAINING SUMMARY OVER SEEDS 42, 2026, AND 3407. VALUES ARE MEAN $\pm$ SAMPLE STANDARD DEVIATION; LOWER FN IS BETTER

Model	mAP@0.5	FN	Fixed Recall	Fixed Precision
YOLO11n	0.8814 $\pm$ 0.0100	205.3 $\pm$ 15.9	0.8846 $\pm$ 0.0090	0.8716 $\pm$ 0.0214
YOLO11n + CGOA	0.8738 $\pm$ 0.0234	206.0 $\pm$ 31.8	0.8843 $\pm$ 0.0179	0.8818 $\pm$ 0.0135
YOLO11n + LSAFF	0.8891 $\pm$ 0.0102	186.0 $\pm$ 15.1	0.8955 $\pm$ 0.0085	0.8681 $\pm$ 0.0214
YOLO11n + CGOA + LSAFF	0.8899 $\pm$ 0.0114	172.7 $\pm$ 19.4	0.9030 $\pm$ 0.0109	0.8662 $\pm$ 0.0209

Fig. 8 places the mean fixed operating points in precision–recall space. The full model occupies the highest-recall position, while CGOA has the highest mean fixed precision but essentially the same mean fixed recall as YOLO11n. LSAFF lies between these outcomes. This view clarifies why the full model is described as recall-oriented rather than uniformly superior.

D. Supplemental Second-Benchmark Evaluation on PIDray

Table X reports the combined Easy, Hard, and Hidden results for the single-seed PIDray experiment. The full model improves standard precision from 0.7832 to 0.7885, standard

recall from 0.7498 to 0.7674, mAP@0.5 from 0.7938 to 0.8035, and mAP@0.5:0.95 from 0.6582 to 0.6698. At the fixed operating point, it detects 830 additional ground-truth instances and reduces FN from 10,148 to 9,318, an 8.18% reduction. The gain is accompanied by 409 additional FP, a 4.36% increase, and a small decrease in fixed precision from 0.7726 to 0.7696. The corresponding descriptive yield is 2.03 additional true detections per additional false positive. Fixed recall increases from 0.7584 to 0.7781 and fixed F_2 from 0.7612 to 0.7764. As with OPIXray, this ratio summarizes the observed count change rather than defining a universal screening utility. The result reproduces the recall-oriented

TABLE IX. INDIVIDUAL RESULTS FOR THE THREE SPLIT-AND-TRAINING TRIALS. FIXED R AND FIXED P USE CONFIDENCE 0.25 AND IoU 0.50

Model	Seed	mAP@0.5	FN	Δ FN	Fixed R	Δ R	Fixed P
YOLO11n	42	0.8852	201	–	0.8871	–	0.8758
YOLO11n	2026	0.8701	223	–	0.8747	–	0.8485
YOLO11n	3407	0.8889	192	–	0.8921	–	0.8906
YOLO11n + CGOA	42	0.8467	240	+39	0.8652	-0.0219	0.8661
YOLO11n + CGOA	2026	0.8879	201	-22	0.8871	+0.0124	0.8896
YOLO11n + CGOA	3407	0.8867	177	-15	0.9006	+0.0085	0.8896
YOLO11n + LSAFF	42	0.8927	184	-17	0.8966	+0.0095	0.8622
YOLO11n + LSAFF	2026	0.8776	202	-21	0.8865	+0.0118	0.8502
YOLO11n + LSAFF	3407	0.8971	172	-20	0.9034	+0.0113	0.8918
YOLO11n + CGOA + LSAFF	42	0.8873	163	-38	0.9084	+0.0213	0.8661
YOLO11n + CGOA + LSAFF	2026	0.8801	195	-28	0.8904	+0.0157	0.8453
YOLO11n + CGOA + LSAFF	3407	0.9023	160	-32	0.9101	+0.0180	0.8872

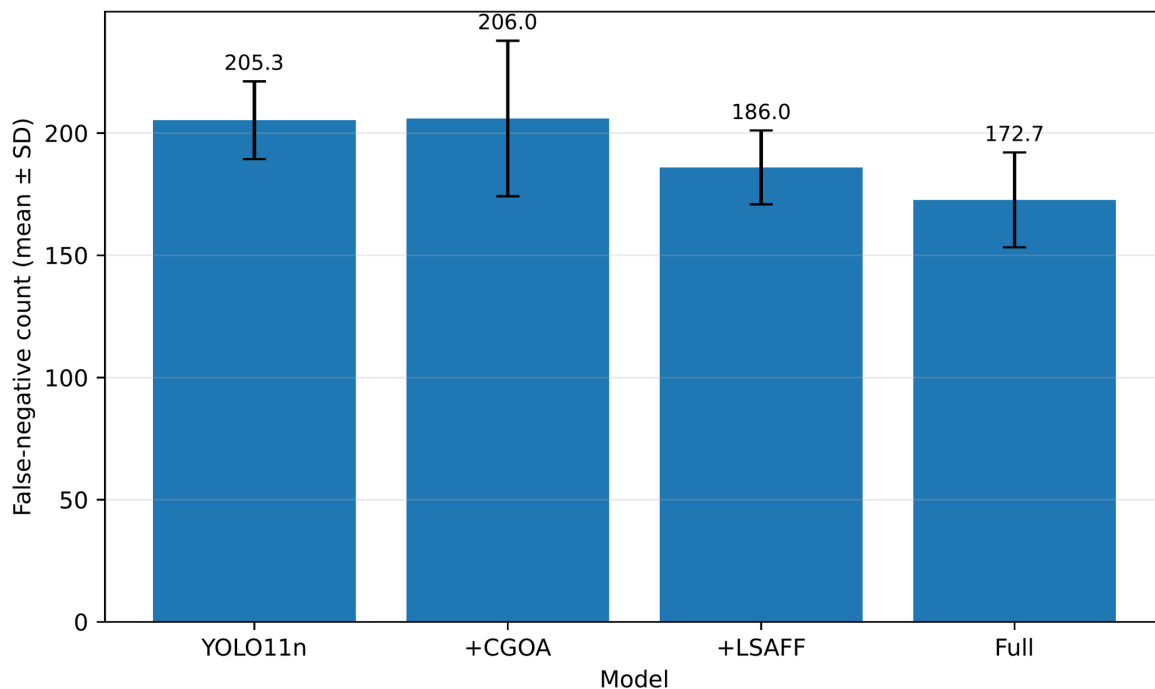


Fig. 7. Mean false-negative count and sample standard deviation over three split-and-training trials, in which both the validation split and training seed change. Lower values are better. The figure summarizes variability and is not a fixed-split significance test.

pattern observed on OPIXray rather than showing a uniform improvement in every error measure.

LSAFF alone already improves the combined PIDray result while keeping fixed precision nearly unchanged (0.7731 versus 0.7726). Adding CGOA within the full model reduces FN by a further 314 relative to LSAFF and gives the best standard recall, mAP, fixed recall, and fixed F_2 , although FP increases and fixed precision decreases. This result is consistent with the OPIXray interpretation that LSAFF is the stronger standalone contributor and CGOA is most useful when coupled with multi-scale fusion.

Fig. 9 separates the three official PIDray difficulty subsets. Relative to YOLO11n, the full model raises mAP@0.5 by 0.0075 on Easy, 0.0080 on Hard, and 0.0135 on Hidden. Fixed recall increases by 0.0174, 0.0149, and 0.0245, respectively, while FN decreases by 147, 216, and 467. The largest absolute FN reduction and the largest fixed-recall improvement occur on Hidden, supporting consistent behavior on a second benchmark that includes deliberately concealed items. Because this experiment uses one split and training seed, the subset differences are descriptive and do not establish statistical significance.

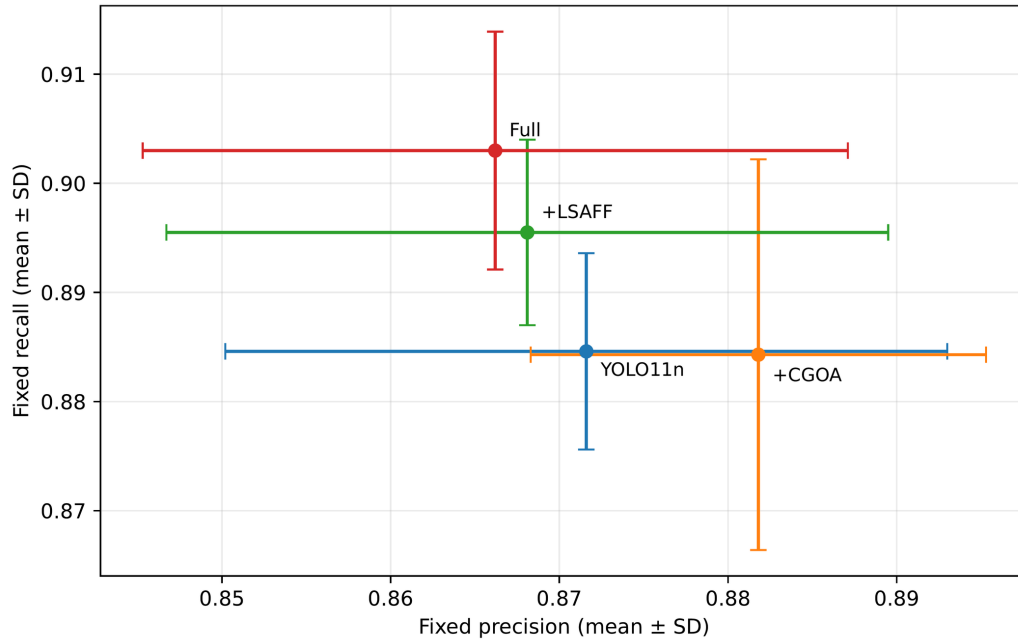


Fig. 8. Mean fixed precision and fixed recall over three split-and-training trials. Horizontal and vertical error bars show sample standard deviations.

TABLE X. SINGLE-SEED SUPPLEMENTAL SECOND-BENCHMARK VALIDATION ON THE COMBINED PIDRAY EASY, HARD, AND HIDDEN TEST SUBSETS. STANDARD P AND R ARE CLASS-AVERAGED VALIDATOR METRICS; FIXED P, FIXED R, AND F_2 ARE MICRO-AVERAGED AT CONFIDENCE 0.25 AND IOU 0.50

Model	P	R	mAP@0.5	mAP@0.5:0.95	TP	FP	FN	Fixed P	Fixed R	F_2
YOLO11n	0.7832	0.7498	0.7938	0.6582	31,852	9,373	10,148	0.7726	0.7584	0.7612
YOLO11n + LSAFF	0.7858	0.7621	0.8012	0.6675	32,368	9,502	9,632	0.7731	0.7707	0.7712
XraySafe-YOLO	0.7885	0.7674	0.8035	0.6698	32,682	9,782	9,318	0.7696	0.7781	0.7764

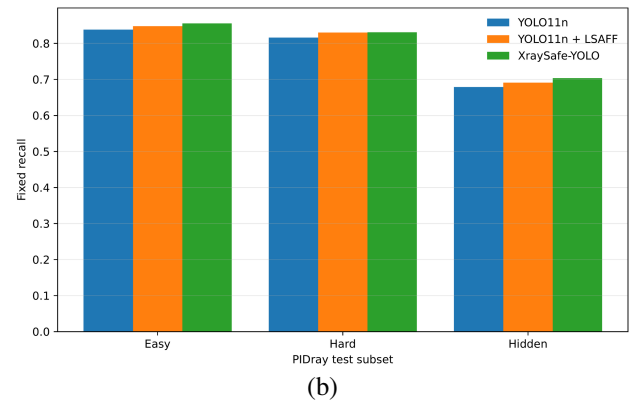
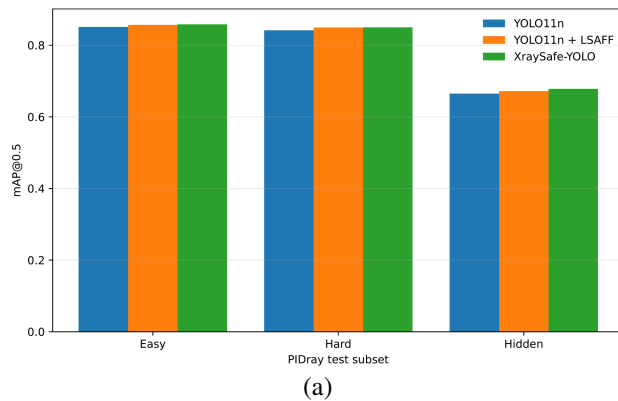


Fig. 9. Single-seed PIDray results across the official Easy, Hard, and Hidden subsets: (a) mAP@0.5 and (b) fixed recall at confidence 0.25 and IoU 0.50. The three subsets contain different numbers of ground-truth boxes; the normalized metrics are therefore more suitable than raw FN counts for cross-subset visual comparison.

E. Confidence Sweep and Operating-Point Interpretation

To test whether the recall advantage is only an artifact of applying the same confidence threshold to differently calibrated models, the saved seed-42 checkpoints were re-evaluated at confidence thresholds 0.10, 0.20, 0.25, 0.30, 0.40, and 0.50. Fig. 10 plots fixed recall against false positives per

image for YOLO11n, LSAFF, and the full model. The full model remains above the baseline over most of the measured false-positive range, while LSAFF is intermediate. This post-hoc sweep uses the test set only to assess operating-point fairness; it is not used to choose a recommended deployment threshold. Any operational threshold should be selected on validation data from the target scanner and workload.

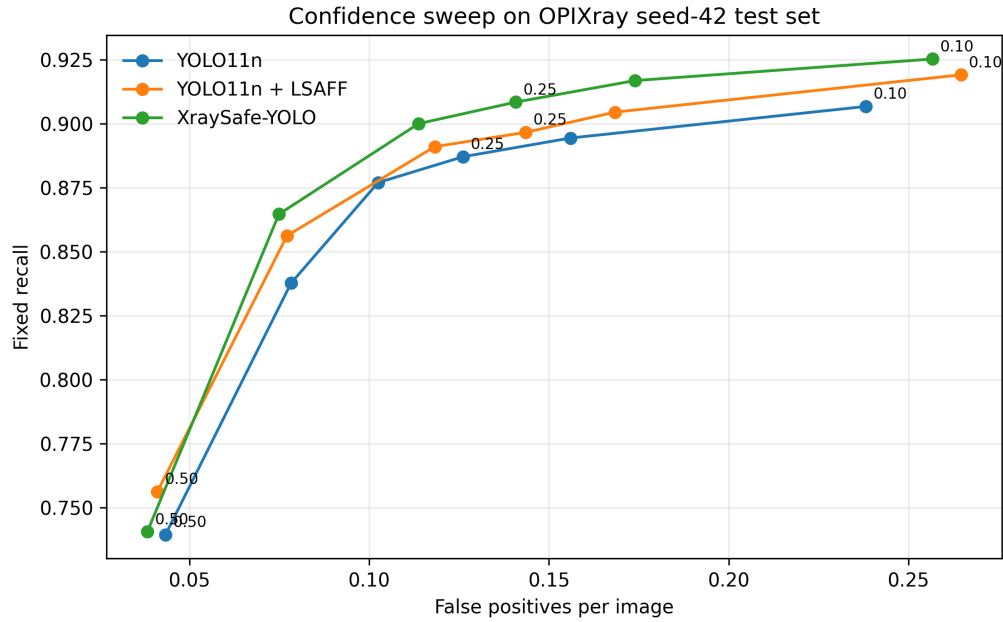


Fig. 10. Seed-42 OPIXray confidence sweep for YOLO11n, YOLO11n + LSAFF, and the full model. Recall is plotted against false positives per image, which avoids assuming that identical confidence thresholds imply equivalent operating conditions. Labels mark selected confidence thresholds.

TABLE XI. SEED-42 OPIXRAY RECALL AT MATCHED FALSE-POSITIVE BUDGETS, OBTAINED BY LINEAR INTERPOLATION BETWEEN ADJACENT MEASURED CONFIDENCE-SWEEP POINTS

Model	150 FP (0.0845 FP/image)	200 FP (0.1126 FP/image)	250 FP (0.1408 FP/image)
YOLO11n	0.8477	0.8813	0.8907
YOLO11n + LSAFF	0.8624	0.8862	0.8960
XraySafe-YOLO	0.8733	0.8990	0.9084

TABLE XII. MODEL COMPLEXITY AT 640-PIXEL INPUT RESOLUTION

Model	Params (M)	GFLOPs	Model Size (MB)
YOLO11n	2.583	6.317	5.216
YOLO11n + CBAM-lite	2.587	6.324	5.231
YOLO11n + CGOA	2.589	6.339	5.238
YOLO11n + LSAFF	2.893	7.227	5.877
YOLO11n + CGOA + LSAFF	2.899	7.249	5.898

For a matched-false-positive view, recall was linearly interpolated between adjacent measured sweep points at common FP budgets of 150, 200, and 250 detections over the 1776-image OPIXray test set. Table XI shows that the full model has the highest recall at all three budgets. At 200 FP, for example, the interpolated recalls are 0.8813 for YOLO11n, 0.8862 for LSAFF, and 0.8990 for the full model. Because these values are interpolated summaries of the measured sweep rather than independently optimized thresholds, they are used only as descriptive matched-operating-point evidence.

The fixed-count comparison at confidence 0.25 and IoU 0.50 remains useful as a common diagnostic point. At this threshold, the combined model reduces FN and increases FP relative to YOLO11n on both OPIXray and PIDray. The OPIXray three-trial trade-off in Fig. 8 indicates that this behavior is not limited to the seed-42 count table: the full

model has the highest mean fixed recall and lower mean fixed precision than the baseline. The single-seed PIDray result shows the same direction on a broader category set. The sweep and matched-FP analysis strengthen the interpretation that the architectural change provides recall benefit beyond simply lowering the baseline detector's confidence threshold, while the final operating threshold remains an application-specific validation choice.

F. Model Complexity and Qualitative Examples

Table XII reports parameter count, GFLOPs, and model size at 640-pixel input resolution. The combined model has 2.899 M parameters and 7.249 GFLOPs, about 12.2% more parameters and 14.8% more computation than YOLO11n. Latency and FPS are not compared because they depend strongly on the local hardware and software environment.

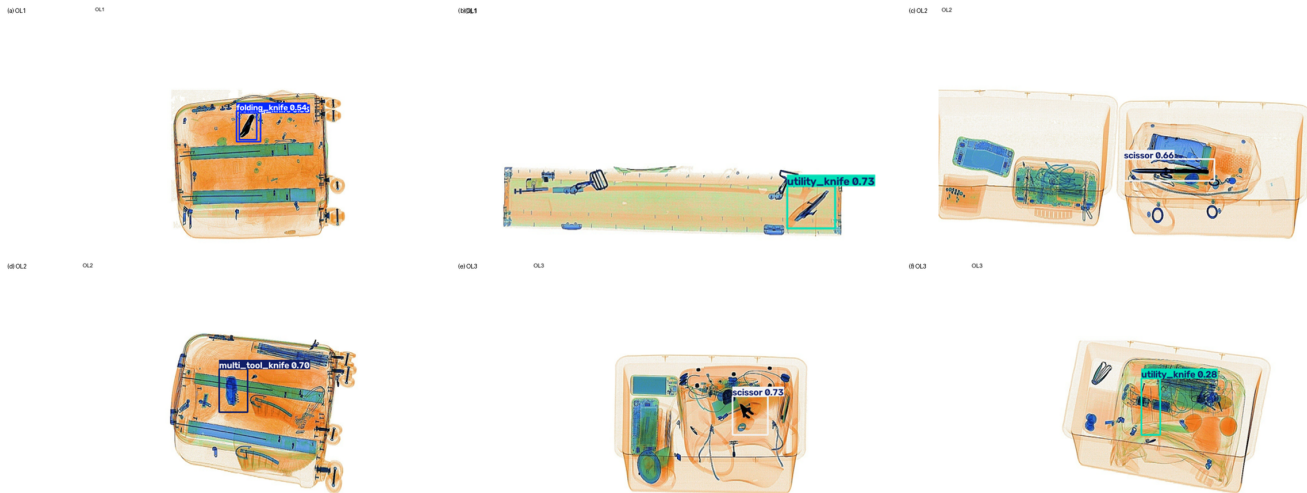


Fig. 11. Representative seed-42 predictions from the combined model: (a,b) OL1, (c,d) OL2, and (e,f) OL3.

Fig. 11 presents six combined-model predictions spanning OL1–OL3. Panels (a) and (b) show compact and elongated cutters whose visible boundaries remain distinguishable from the surrounding baggage structure. Panels (c) and (d) illustrate localization when a prohibited item overlaps dense electronics or elongated metal objects. Panels (e) and (f) are drawn from OL3; the utility-knife prediction in panel (f) has confidence 0.28, close to the fixed-count threshold, and illustrates the narrow margin available in a heavily cluttered scene. The examples show that the model can produce usable localizations at each occlusion level, but they are not a paired baseline comparison. Relative conclusions therefore rely on the complete test-set metrics and error counts rather than on selected images.

VI. DISCUSSION

A. What the Experiments Show

The two modules have different effects. LSAFF is the more stable individual module. Across the three trials, it improves mean mAP@0.5, reduces mean FN, and increases mean fixed recall. This is reasonable for OPIXray because the detector needs both high-resolution details from partly visible blades and wider context from nearby objects. LSAFF lets each detection scale use information from P3, P4, and P5.

CGOA is less stable. Boundary-motivated local reweighting may help a partly visible item, but X-ray baggage also contains many strong edges from wires, frames, and overlapping materials. The CGOA model is below the baseline at seed 42 and has the largest variation across the three trials. Its mean FN is almost the same as YOLO11n. When the CGOA-enhanced P3 and P4 features are fused with P5 by LSAFF, the full model gives a lower mean FN than LSAFF alone. These results suggest that CGOA is useful mainly as a complement to multi-scale fusion, not as a better standalone attention module.

The class and occlusion results also show that the gain is not uniform. Multi-tool and utility knives improve, folding-knife recall decreases slightly, and straight knife remains difficult. The full model has the lowest FN on OL1 and OL3, while LSAFF is best on OL2. The clearest benefit of the

full model is therefore fewer missed detections at the selected threshold.

The PIDray result shows a similar pattern on 12 classes. LSAFF improves all reported aggregate metrics and keeps fixed precision close to the baseline. The full model further reduces FN on Easy, Hard, and Hidden, with the largest fixed-recall gain on Hidden. FP also increases, so the PIDray result supports the same recall-oriented trade-off rather than a general improvement in every metric.

B. Relation to Recent X-ray YOLO Research

Recent compact X-ray detectors often combine several improvements and test them on more than one dataset. Examples include channel reduction with BiFPN [15], feature-aggregation redesign with normalized Wasserstein distance [16], state-space blocks evaluated on three datasets [21], contrastive multi-scale learning [22], and sparsity with adaptive distillation [23]. Those studies provide broader evidence on multiple runs, datasets, or inference speed than this study.

XraySafe-YOLO uses repeated OPIXray ablations and a single-seed PIDray second-benchmark test. It keeps the YOLO11n baseline small, separates the proposed modules, and exposes the fixed-threshold TP/FP/FN trade-off on both benchmarks. This is useful because a higher mAP alone does not show how many missed items are recovered at a selected threshold. At the same time, the PIDray result is not a state-of-the-art benchmark study and should not be compared directly with models trained under different splits, schedules, resolutions, or augmentation protocols.

C. Operational Meaning of the FP–FN Trade-off

At confidence 0.25, the seed-42 OPIXray full model finds 38 additional ground-truth objects and produces 26 additional false positives relative to YOLO11n, equivalent to 1.46 recovered objects per added false alarm. Across all three OPIXray trials, it lowers FN and raises fixed recall relative to the corresponding YOLO11n run, while fixed precision remains

slightly lower. PIDray shows the same direction: the full model recovers 830 additional ground-truth instances and produces 409 additional false positives, or 2.03 recovered objects per added false alarm. These ratios only describe the observed counts. A missed dangerous item and an unnecessary review do not have the same operational cost. In a human-in-the-loop screening system, the observed shift may be reasonable when the cost of a missed prohibited item is greater than the cost of secondary inspection. It is not automatically preferable, however. Additional false alarms consume operator time, can slow baggage throughput, and may reduce trust if the workload becomes excessive.

The preferred threshold depends on staffing, scanner type, traffic volume, and the relative cost of missed items and false alarms. The fixed threshold in this study is a common diagnostic point. Before deployment, threshold selection should use data from the target scanner, include calibration and confidence-sweep analysis, and report recall at a matched false-positive rate or operational workload.

D. Efficiency and Practical Use

The full model remains compact, with 2.899 M parameters and 7.249 GFLOPs at an input size of 640 pixels. Its parameter count is 12.2% above YOLO11n and its theoretical computation is 14.8% higher. These results show that the added modules have limited structural overhead. They do not prove real-time speed or hardware efficiency. FLOPs do not replace measured latency, and recent X-ray studies that use deployment-oriented language also report FPS or hardware-specific inference time [21], [23]. Accordingly, the present study uses “compact” for the detector as a whole and uses “lightweight” only for individual low-overhead operations or modules, and when describing related work that adopts that design objective.

The detector is intended to mark candidate regions for an inspector. It is not designed to make an automatic release decision. The operating threshold, escalation rules, and review interface would need to be evaluated together with detector accuracy in an operational study.

E. Limitations and Future Work

The primary repeated-trial evidence remains limited to OPIXray and its five cutter-like categories. PIDray adds a broader 12-class second-benchmark test with official Easy, Hard, and Hidden subsets, but it uses only one validation split and training seed and compares three configurations rather than the complete ablation. It therefore provides useful second-benchmark evidence without establishing robustness across repeated PIDray runs, HiXray, CLCXray, dual-view systems, other scanner manufacturers, cargo images, or operational customs data. Several recent models report results on two or three datasets [15], [21]–[23], so broader cross-domain evaluation remains important.

The repeated evaluation includes only three trials and changes the validation split together with the training seed. The reported standard deviations therefore combine split and optimization variability and are not fixed-split significance estimates. The overlapping error bars in Fig. 7 and 8 should be

interpreted accordingly. A stronger design would keep the validation set fixed, repeat every configuration with more seeds, and apply paired bootstrap analysis to saved test predictions.

Other limitations concern experimental coverage. The automatic optimizer policy can couple architecture choice with automatically selected optimizer settings. The generic attention comparison is limited to CBAM-lite, and ASFF or BiFPN were not trained under the same protocol. CGOA alone was not repeated on PIDray. No fixed-split significance test, image-level bootstrap interval, measured latency, direct fusion-baseline training, attention-map analysis, or paired baseline qualitative panel is reported. For this reason, the manuscript does not claim real-time performance. CGOA uses no contour supervision and remains unstable as a standalone module on OPIXray. Very small, low-contrast, heavily overlapped, and visually ambiguous objects also remain difficult.

Future work will prioritize multi-seed second-benchmark validation on additional datasets, fixed-split repeated training, direct fusion baselines, confidence calibration, measured inference speed, and paired visual error analysis. A human-in-the-loop study should then quantify whether the reduction in missed detections justifies the additional secondary-inspection workload.

VII. CONCLUSIONS

XraySafe-YOLO combines scale-weighted fusion and contour-motivated attention in a compact YOLO11n detector. On OPIXray, LSAFF is the stronger standalone module, while CGOA is more sensitive to the validation split and training seed. In all three trials, the full model reduces FN and increases fixed recall compared with YOLO11n. It has the lowest mean false-negative count (172.7 ± 19.4) and the highest mean fixed recall (0.9030 ± 0.0109). At seed 42, FN decreases from 201 to 163 and fixed F_2 increases from 0.8848 to 0.8996, while FP increases from 224 to 250. The class and occlusion results show that the improvement is useful but not the same for every case.

The PIDray experiment shows the same main direction on 12 classes. mAP@0.5 increases from 0.7938 to 0.8035, FN decreases from 10,148 to 9,318, and fixed recall increases from 0.7584 to 0.7781. Fixed precision decreases slightly from 0.7726 to 0.7696. The largest fixed-recall gain is on the Hidden subset. Because PIDray is tested with one seed, this result supports the architecture on a second dataset but does not replace repeated second-benchmark validation.

The model should be viewed as a compact recall-oriented option for human-assisted X-ray screening, not as a detector that is better in every metric or ready for direct deployment. The main evidence comes from the fixed-threshold error analysis, repeated OPIXray trials, and the PIDray second-benchmark test. Stronger conclusions would require repeated tests on more datasets and scanners, direct fusion baselines, threshold calibration, and measured inference speed.

ACKNOWLEDGMENT

The authors thank the OPIXray and PIDray maintainers for making the benchmarks available to the research community.

DECLARATIONS

Funding: This research was funded by the Scientific Research Startup Fund of Shanghai Customs College under Grant No. 2315053A2024.

Data availability: The experiments use the public OPIXray and PIDray benchmarks. The original images and annotations are not redistributed and should be obtained from the official repositories at <https://github.com/OPIXray-author/OPIXray> and <https://github.com/lutao2021/PIDray>, subject to their academic-use terms. Processed result summaries, model-structure code, configuration files, and demonstration materials supporting the findings are available from the corresponding author upon reasonable request.

Conflict of interest: The authors declare no conflict of interest.

REFERENCES

- [1] C. Miao, L. Xie, F. Wan, S. Su, H. Liu, J. Jiao, and Q. Ye, "SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2119–2128.
- [2] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An X-ray security inspection benchmark and de-occlusion attention module," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 138–146, doi: 10.1145/3394171.3413828.
- [3] OPIXray Authors, "OPIXray: Occluded prohibited items X-ray dataset." [Online]. Available: <https://github.com/OPIXray-author/OPIXray>. Accessed: Aug. 18, 2026.
- [4] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards real-world prohibited item detection: A large-scale X-ray benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 5412–5421.
- [5] L. Zhang, L. Jiang, R. Ji, and H. Fan, "PIDray: A large-scale X-ray benchmark for real-world prohibited item detection," *Int. J. Comput. Vis.*, vol. 131, pp. 3170–3192, 2023, doi: 10.1007/s11263-023-01855-1.
- [6] L. Zhang, L. Jiang, R. Ji, and H. Fan, "PIDray: A large-scale X-ray benchmark for real-world prohibited item detection." [Online]. Available: <https://github.com/lutao2021/PIDray>. Accessed: Aug. 18, 2026.
- [7] M. Li, T. Jia, H. Wang, B. Ma, H. Lu, S. Lin, D. Cai, and D. Chen, "AO-DETR: Anti-overlapping DETR for X-ray prohibited items detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, pp. 12076–12090, 2025, doi: 10.1109/TNNLS.2024.3487833.
- [8] R. Tao, H. Wang, Y. Guo, H. Chen, L. Zhang, X. Liu, Y. Wei, and Y. Zhao, "Dual-view X-ray detection: Can AI detect prohibited items from dual-view X-ray images like humans?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 10338–10347.
- [9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.
- [10] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [11] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [12] Ultralytics, "YOLOv8 documentation." [Online]. Available: <https://docs.ultralytics.com/>. Accessed: Aug. 18, 2026.
- [13] Ultralytics, "YOLO11 documentation." [Online]. Available: <https://docs.ultralytics.com/models/yolo11/>. Accessed: Aug. 18, 2026.
- [14] X. Huang and Y. Zhang, "ScanGuard-YOLO: Enhancing X-ray prohibited item detection with significant performance gains," *Sensors*, vol. 24, Art. no. 102, 2024, doi: 10.3390/s24010102.
- [15] Y. Fang, C. Xu, and Y. Zhang, "Research on X-ray security contraband identification technology based on lightweight YOLOv8," *Sci. Rep.*, vol. 14, Art. no. 25031, 2024, doi: 10.1038/s41598-024-75932-x.
- [16] Y.-T. Zhou, K.-Y. Cao, D. Li, and J.-C. Piao, "Fine-YOLO: A simplified X-ray prohibited object detection network based on feature aggregation and normalized Wasserstein distance," *Sensors*, vol. 24, Art. no. 3588, 2024, doi: 10.3390/s24113588.
- [17] H. Zhang, Z. Zhao, and J. Yang, "Attention-based prohibited item detection in X-ray images during security checking," *IET Image Process.*, vol. 18, pp. 1119–1131, 2024, doi: 10.1049/ipr2.13013.
- [18] H. Zhang, W. Teng, X. He, H. Que, and Y. Zhang, "Lightweight prohibited items detection model in X-ray images based on improved YOLOv7-Tiny," *J. Franklin Inst.*, vol. 362, Art. no. 107421, 2025, doi: 10.1016/j.jfranklin.2024.107421.
- [19] F. Guan, H. Zhang, and X. Wang, "An improved YOLOv8 model for prohibited item detection with deformable convolution and dynamic head," *J. Real-Time Image Process.*, vol. 22, Art. no. 84, 2025, doi: 10.1007/s11554-025-01665-3.
- [20] J. Chen, J. Hao, and X. Liu, "An X-ray contraband detection method based on improved YOLOv8," *IET Image Process.*, vol. 19, Art. no. e70135, 2025, doi: 10.1049/ipr2.70135.
- [21] K. Zhao, S. Peng, Y. Li, and T. Lu, "A lightweight Xray-YOLO-Mamba model for prohibited item detection in X-ray images using selective state space models," *Sci. Rep.*, vol. 15, Art. no. 13171, 2025, doi: 10.1038/s41598-025-96035-1.
- [22] Q. Diao, W. Chan, A. M. Zain, K. Moorthy, Y. Chen, H. Tong, and Y. Huo, "A lightweight multi-scale detection framework for X-ray images with supervised contrastive learning," *Sci. Rep.*, vol. 16, Art. no. 8635, 2026, doi: 10.1038/s41598-026-38000-0.
- [23] J. Kong, "GSA-YOLO enables high-efficiency real-time X-ray security inspection via structured sparsity and adaptive knowledge distillation," *Sci. Rep.*, advance online publication, 2026, doi: 10.1038/s41598-026-55599-2.
- [24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520.
- [26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [27] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13713–13722.
- [28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.
- [29] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8759–8768.
- [30] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10781–10790.
- [31] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," *arXiv preprint arXiv:1911.09516*, 2019.
- [32] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21002–21012.