buildings

MDPI

Article

HFE-YOLO: Hybrid Feature Enhancement with Multi-Attention Mechanisms for Construction Site Object Detection

Teerapun Saeheaw

Department of Teacher Training in Mechanical Engineering, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand; teerapun.s@fte.kmutnb.ac.th

Abstract

Construction sites require integrated monitoring of equipment and structural safety. This study systematically compares four feature enhancement mechanisms at the YOLOv11n backbone–neck transition: HFE-YOLO (hybrid multi-attention), FPN-YOLO (feature pyramid), C2F-YOLO (cross-stage partial), and Identity-YOLO (baseline). Evaluation utilized two datasets with contrasting class distributions: ConstructSight (eight equipment classes, severe imbalance) and SafeGuard (five safety classes, balanced distribution). All models were trained for 200 epochs using identical configurations to ensure controlled comparison. On the imbalanced CS dataset, HFE-YOLO achieves superior performance (95.0% mAP@50, 82.6% mAP@50–95), followed by FPN-YOLO (94.8%, 82.4%), Identity-YOLO (92.5%, 74.4%), and C2F-YOLO (92.4%, 72.1%). On the balanced SG dataset, performance differences compress substantially: HFE-YOLO (96.8%, 79.4%), C2F-YOLO (96.6%, 78.2%), Identity-YOLO (96.3%, 78.1%), and FPN-YOLO (96.1%, 76.1%). HFE-YOLO provides 8.2 percentage points mAP@50–95 improvement over the baseline on imbalanced data versus 1.3 percentage points on balanced data. Enhancement mechanism effectiveness varies substantially between dataset distributions, with sophisticated mechanisms providing greater benefits for imbalanced scenarios. These findings offer insights for architecture selection based on dataset distribution characteristics.

Keywords: feature enhancement; YOLOv11; multi-attention; class imbalance; construction monitoring; object detection

check for updates

Academic Editor: Antonio Caggiano

Received: 24 October 2025
Revised: 21 November 2025
Accepted: 24 November 2025
Published: 26 November 2025

Citation: Saeheaw, T. HFE-YOLO: Hybrid Feature Enhancement with Multi-Attention Mechanisms for Construction Site Object Detection. *Buildings* **2025**, *15*, 4274. https://doi.org/10.3390/buildings15234274

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Construction sites represent complex operational environments where the simultaneous monitoring of heavy equipment and structural safety compliance presents significant challenges for workplace safety management. The construction industry faces persistent safety challenges, with equipment-related incidents and structural safety violations representing significant risk factors for workplace accidents. Heavy equipment-related safety accidents account for a significant proportion of construction fatalities, necessitating more effective monitoring approaches [1].

Traditional monitoring methods, including manual inspections and conventional video surveillance, often lack real-time performance and comprehensive coverage, making them inefficient for effective site supervision [2,3].

Recent advances in computer vision have shown measurable progress in automated construction site monitoring, with deep learning approaches achieving detection accuracies ranging from 73.2 to 98.93% mAP@0.5—including 98.93% for heavy equipment detection [1], 97.9% for scaffolding safety compliance [4], and 97.03% for machinery swarm

operations [5]—establishing the viability of automated monitoring systems (Table 1). Object detection architectures, particularly the YOLO series and transformer-based approaches, have demonstrated significant capabilities in detecting construction equipment and monitoring safety compliance elements in real-time [6,7].

Construction monitoring research demonstrates diverse methodological approaches addressing different aspects of the monitoring challenge. Algorithmic optimization studies focus on architecture enhancement and real-time processing capabilities [8–10]. Complementary systems integration research addresses end-to-end monitoring workflows through Building Information Modeling (BIM) integration with computer vision for automated progress monitoring [11], regulatory compliance frameworks for automated monitoring deployment addressing privacy and workplace surveillance requirements [12,13], and digital twin approaches combining wearable sensors with 4D BIM for real-time risk monitoring [14]. These methodological approaches provide comprehensive foundation for construction monitoring system development, combining algorithmic performance optimization with practical deployment frameworks.

Advanced architectural developments have shown promising results for complex detection scenarios. Attention mechanisms have shown effectiveness in refining spatial feature encoding for construction scenarios [15,16]. Feature Pyramid Network (FPN) architectures enable multi-scale processing capabilities essential for detecting objects of varying sizes within construction environments [17,18]. Cross-Stage Partial (CSP) connections have shown potential in optimizing gradient flow and computational efficiency in deep detection networks [19].

Automated construction monitoring deployment operates within regulatory frameworks governing workplace surveillance, data protection, and occupational safety that significantly influence research methodologies. The General Data Protection Regulation (GDPR) Article 88 establishes requirements for processing employee data in employment contexts, mandating Member State rules for 'monitoring systems at the workplace' to safeguard worker dignity and fundamental rights [13–20]. Framework Directive 89/391/EEC establishes occupational safety obligations requiring systematic risk assessments and monitoring systems in alignment with prevention principles [21,22]. Construction Sites Directive 92/57/EEC implements these principles for temporary or mobile construction sites, mandating safety coordinators, health and safety plans, and coordination of prevention measures across contractors [23]. The EU Artificial Intelligence Act [24] classifies AI systems for worker management as high-risk applications requiring conformity assessments, technical documentation, and transparency provisions.

These regulatory requirements fundamentally shape both legal and technical research approaches. Legal research focuses on compliance frameworks, privacy protection mechanisms, and regulatory interpretation across jurisdictions [12,13]. Technical research must address system transparency, explainable AI requirements, and documented performance validation essential for regulatory compliance [11,14]. The present study's controlled architectural comparison methodology directly addresses these dual requirements by providing systematic performance documentation and transparent enhancement mechanism evaluation, establishing an empirical foundation for both technical optimization and regulatory-compliant deployment strategies.

Despite these technological advances, systematic analysis of seventeen recent construction monitoring studies reveals important limitations in addressing integrated equipment detection and structural safety monitoring as unified challenges (Table 1). Current approaches face three primary limitations: methodological constraints, equipment coverage gaps, and safety monitoring integration challenges.

Methodological constraints emerge in architectural evaluation approaches. Prior studies predominantly implement enhancement mechanisms within complete architectural modifications without isolating specific component contributions. For instance, recent work employs efficient channel attention modules [25], small object detection mechanisms [9], or multi-task frameworks [26] alongside multiple architectural modifications, making it difficult to attribute performance improvements to specific enhancement components versus base architecture changes. He et al. [27] employ YOLOv11x-Seg with C3K2 modules and C2PSA attention for segmentation tasks, while Zhang et al. [10] implement MSP-YOLO based on YOLOv12n with multiple enhancement mechanisms. These studies typically modify multiple architectural components simultaneously, precluding controlled assessment of individual enhancement contributions. Furthermore, architectural evaluations predominantly employ single datasets with specific distribution characteristics, limiting understanding of how enhancement effectiveness varies across different class distribution scenarios.

Equipment detection research has concentrated predominantly on commonly visible equipment such as excavators and dump trucks, systematically underrepresenting specialized machinery including diesel generators, tower hoist lifts, and tractors essential for comprehensive site coverage. Analysis of major datasets reveals this pattern: MOCS covers thirteen equipment types but achieves only 51.04% mAP across 41,668 images [28], while ACID addresses ten machinery classes across 10,000 images with 89.2% mAP [8]. Recent specialized studies further narrow this scope: Shin et al. [1] evaluate nine equipment types achieving 98.93% mAP across 10,294 images, while Eum et al. [29] address only five heavy equipment classes achieving 89.72% mAP across 21,772 images. These approaches typically address limited equipment types without incorporating structural safety elements [27,30].

The most significant gap emerges in structural safety compliance monitoring, where major construction monitoring datasets provide extensive equipment coverage yet systematically exclude structural safety features. The SODA dataset includes scaffold as a material category within comprehensive 4M1E framework coverage across 19,846 images, yet lacks safety compliance specifications for guardrail presence or outrigger deployment despite systematic multi-site data collection [31]. MOCS dataset encompasses 13 moving object categories across 174 construction sites and 41,668 images, mentioning scaffold work in documentation but providing no scaffolding safety monitoring classes [28]. Among seventeen recent studies (2021–2025), only Abbas et al. [4] explicitly addresses scaffolding structural safety through guardrail detection and outrigger configuration verification across 4868 images, achieving 97.9% mAP yet representing isolated research without integration with comprehensive equipment detection systems.

Table 1 presents a comparative analysis of representative datasets demonstrating this systematic gap in equipment-safety integration.

Table 1. Comparative analysis of dataset attributes for construction monitoring research.

Dataset/Study	Total Images	Equipment Classes	Safety-Related Classes	Safety Annotation Type	Equipment-Safety Integration
ACID [8]	10,000	10 (machinery)	0	None	No
MOCS [28]	41,668	13 (moving objects)	0	None	No
SODA [31]	19,846	12 (equipment + materials) [a]	3 (worker, helmet, vest)	PPE only	Partial (no structural safety)
[1]	10,294	9 (heavy equipment)	0	None	No
[26]	51,630 [b]	6 (equipment) [c]	2 (workers, helmets)	PPE only	No (multi-task but no structural)

Table 1. *Cont.*

Dataset/Study	Total Images	Equipment Classes	Safety-Related Classes	Safety Annotation Type	Equipment-Safety Integration
[32]	44,295	5 (heavy equipment)	0	None	No
[29]	21,772	5 (heavy equipment)	0	None	No
[4]	4868	0	5 (guardrails, scaffolds, helmets)	Structural + PPE	Safety-only (no equipment)
This Study	14,840 [d]	8 (specialized equipment)	5 (structural + PPE)	Structural + PPE	Yes (unified framework)

Notes: [a] SODA contains 15 total classes: 12 equipment and material categories (board, wood, rebar, brick, scaffold, handcart, cutter, electric box, hopper, hook, fence, slogan) and 3 safety-related classes (worker, helmet, vest). [b] Liu et al. [26] combine three datasets: SODA (19,846 images), MOCS (23,404 images), and Excavator Pose Dataset (8380 images). [c] Liu et al. [26] equipment classes include: excavators, cranes, trucks, loaders, concrete mixers, and additional equipment from combined datasets. Workers and helmets represent safety-related classes. [d] This study combines the ConstructSight dataset (2195 images) and SafeGuard dataset (12,645 images).

Equipment classes include machinery and vehicles; safety-related classes include both PPE and structural safety features (e.g., guardrails, scaffold configurations); equipment-safety integration indicates whether datasets address both categories within a unified framework.

As demonstrated in Table 1, among nine representative studies spanning 2021–2025, no prior work addresses both comprehensive equipment detection and structural safety compliance within a unified framework, with datasets either focusing exclusively on equipment (ACID [8], MOCS [1,28,29,32]: zero safety classes) or limiting safety monitoring to PPE without structural features (SODA [26,31]: PPE only). This systematic exclusion of integrated equipment-safety monitoring represents a fundamental gap in current research approaches.

This comparative investigation addresses three research questions derived from these identified gaps:

RQ1 (Enhancement Mechanism Effectiveness): Do feature enhancement mechanisms at the backbone–neck transition point provide measurable performance improvements, and how does effectiveness vary across different enhancement approaches?

RQ2 (Distribution-Dependent Performance): How does enhancement mechanism effectiveness vary between datasets with contrasting class distribution characteristics?

RQ3 (Integrated Monitoring Capability): Can a unified architectural framework effectively address both equipment detection and structural safety compliance monitoring?

These research questions guide the experimental design and analytical framework presented in subsequent sections.

This study systematically compares four architectural variants at the backbone–neck transition layer of YOLOv11n: HFE-YOLO employing hybrid multi-attention processing through Channel Block Attention Module (CBAM), triplet attention, and spatial attention; FPN-YOLO implementing lightweight FPN architecture; C2F-YOLO utilizing CSP connections; and Identity-YOLO providing a minimal intervention baseline for controlled comparison. Unlike prior construction monitoring studies that evaluate complete architectural modifications or compare different base models (Table 1), this investigation isolates enhancement mechanism effectiveness at the backbone–neck interface while maintaining identical backbone, neck, and head components across all variants, enabling direct quantification of enhancement contributions independent of other architectural variations. This controlled isolation approach contrasts with existing comparative studies that modify multiple architectural components simultaneously [10,25,26], enabling precise attribution of performance improvements to specific enhancement mechanisms. The evaluation utilizes

two specialized datasets—ConstructSight covering eight equipment classes with severe class imbalance, and SafeGuard addressing five structural safety compliance classes with balanced distribution—enabling systematic comparison of architectural effectiveness under distinct distribution scenarios.

2. Theoretical Background and Related Work

2.1. Object Detection Foundations

Modern object detection architectures build upon fundamental computer vision principles established through decades of research. Two-stage detection frameworks, pioneered by Girshick et al. [33] with R-CNN, introduced region proposal mechanisms that separate object localization from classification tasks. This paradigm was later unified in single-stage approaches through YOLO [34] and SSD [35], which directly predict bounding boxes and class probabilities in single network passes.

Feature pyramid architectures represent critical theoretical advancement in handling multi-scale object detection challenges. Lin et al. [17] demonstrated that hierarchical feature representations enable effective detection across object scales by combining semantically strong deep features with spatially precise shallow features. This theoretical framework underlies most contemporary detection architectures including YOLO variants.

Attention mechanisms derive from the transformer architecture principle that different spatial regions contribute unequally to detection performance [36]. Channel attention, formalized by Hu et al. [37] through squeeze-and-excitation networks, enables adaptive feature recalibration by modeling interdependencies between feature channels. Spatial attention complements this by emphasizing informative spatial locations while suppressing irrelevant background regions [38].

2.2. Multi-Task Learning Framework

Multi-task learning theory, established by Caruana [39], demonstrates that related tasks can benefit from shared feature representations when underlying data distributions share common structure. In construction monitoring contexts, equipment detection and safety compliance monitoring represent related visual recognition tasks that may benefit from shared feature representations.

Research in multi-task learning indicates that unified architectures can achieve superior performance compared to separate specialized models when tasks exhibit complementary learning signals, though task interference can occur when learning objectives conflict [40]. This requires careful architecture design to balance shared and task-specific feature processing.

2.3. Construction Monitoring Approaches

Construction monitoring encompasses diverse methodological approaches beyond computer vision. Sensor-based structural health monitoring represents a parallel research domain, with notable advances including partial model-based damage identification using stiffness separation methods for long-span steel truss bridges [41], computational efficiency improvements through stiffness separation approaches for truss structure damage identification [42], and optimization of sensor placement strategies for structural damage detection [43]. These sensor-based methodologies employ physical instrumentation and mathematical modeling for post-construction damage assessment, complementing vision-based approaches that focus on construction activity monitoring and real-time visual detection applications.

3. Methodology

3.1. YOLOv11 Architecture Overview

YOLOv11 incorporates architectural innovations across multiple model sizes including nano (n), small (s), medium (m), large (l), and extra-large (x) variants [44]. The YOLOv11n model was selected as the baseline for this study due to its computational efficiency characteristics suitable for real-time deployment in construction site monitoring [45]. With only 4.7M parameters and 9.3 GFLOPs, YOLOv11n achieves 43.2% mAP@0.5:0.95 on COCO dataset [46], representing a 1.4% mAP improvement over YOLOv10n alongside reduced inference latency [46]. While larger variants (s/m/l/x) offer higher accuracy, their increased computational demands compromise real-time processing capabilities essential for resource-constrained edge device deployment [46].

The backbone employs C3k2 blocks as refined CSP bottlenecks and incorporates SPPF for spatial pyramid pooling, followed by C2PSA for spatial attention enhancement [44]. The neck utilizes C3k2 blocks for multi-scale feature aggregation, while the head processes refined features through detection layers to generate final predictions [44].

3.2. Enhancement Mechanisms

Figure 1a illustrates the complete YOLOv11n architecture employed across all experimental variants, comprising backbone (Layers 0–9), enhancement module (Layer 10), neck (Layers 11–22), and detection head (Layer 23). Layer 10 represents the sole architectural variation point, with all other components remaining identical across variants.

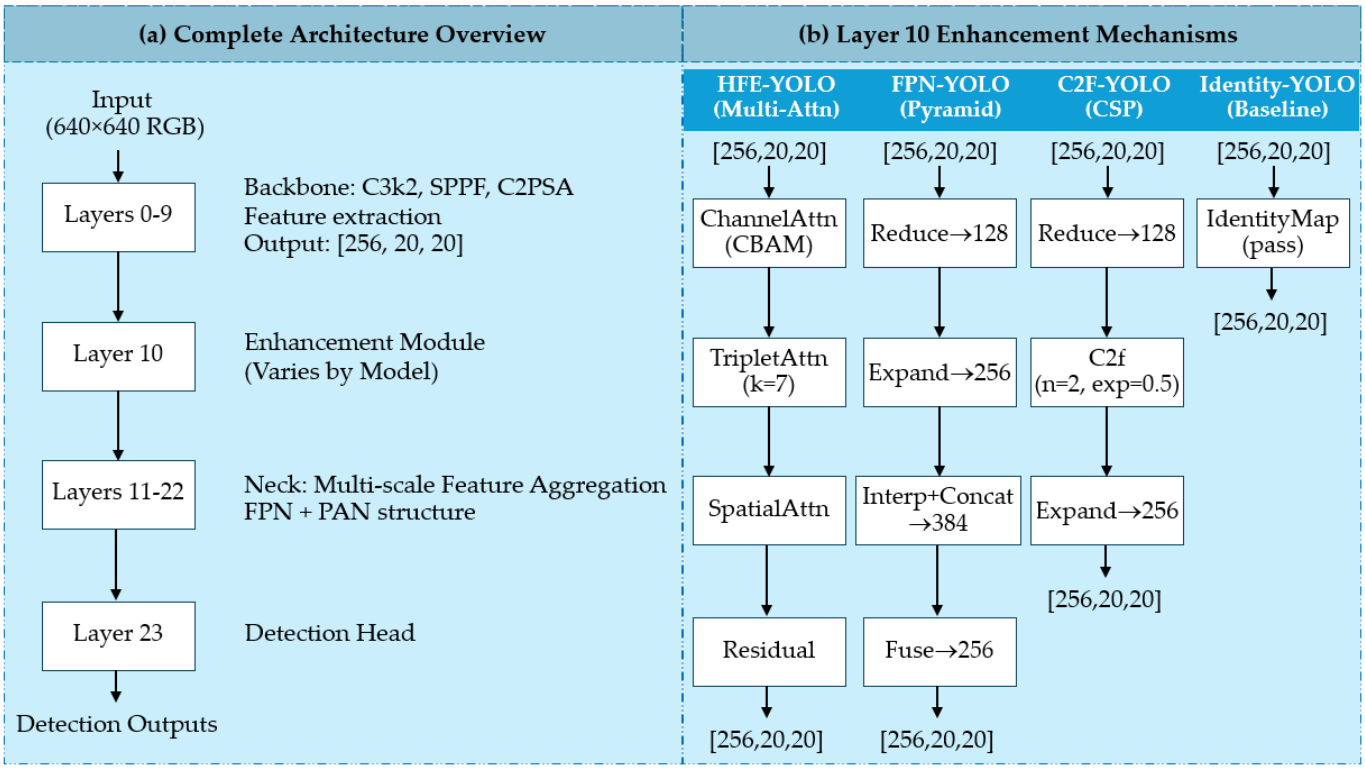


Figure 1. YOLOv11n architecture with Layer 10 enhancement variants.

The backbone–neck transition point (Layer 10) was selected as the enhancement location based on three architectural considerations supported by object detection theory. First, this position represents the critical interface where high-level semantic features from the backbone must be transformed for multi-scale processing in the neck, making it a natural point for feature enhancement operations [17]. Second, the backbone–neck transition

operates on intermediate feature dimensions [256, 20, 20] that balance computational efficiency with representational capacity, enabling effective attention mechanisms without excessive computational overhead compared to earlier high-resolution layers [38]. Third, this location enables controlled comparison of enhancement mechanisms while maintaining consistent feature extraction (Layers 0–9) and multi-scale aggregation pipelines (Layers 11–22) across all experimental variants, isolating the contribution of specific enhancement approaches independent of other architectural variations.

3.2.1. HFE-YOLO: Multi-Attention Processing

The proposed HFE-YOLO architecture introduces a Hybrid Feature Enhancement (HFE) module at Layer 10 of the YOLOv11n backbone, positioned at the backbone–neck transition point. HFE-YOLO operates on feature tensors of shape [batch, 256, 20, 20] output from Layer 9 (SPPF), where 256 represents the number of feature channels encoding semantic information, and 20 × 20 denotes the spatial resolution (height × width) of the feature maps at this hierarchical level for 640 × 640 input images. This intermediate resolution—representing a 32× down sampling from the input—balances computational tractability with sufficient spatial granularity for object localization, as established in feature pyramid literature [17]. The selection of this specific layer stems from its position as the semantic-to-spatial transformation interface: Layer 9 completes high-level semantic feature extraction through spatial pyramid pooling, while subsequent neck layers (11–22) perform multi-scale feature aggregation for detection. The HFE module's multi-attention mechanisms operate on these 256-channel features to refine spatial feature encoding in three dimensions: channel attention recalibrates feature importance across the 256 channels based on global context [38], triplet attention models cross-dimensional dependencies across channel-height-width axes [47], and spatial attention emphasizes informative spatial locations within the 20 × 20 feature maps while suppressing irrelevant background regions. This hierarchical refinement enhances feature discriminability for subsequent multi-scale detection, particularly benefiting minority class detection in imbalanced scenarios where subtle feature distinctions become critical.

As detailed in Figure 1b, the HFE module implements hierarchical attention processing through three complementary mechanisms. The channel attention component captures inter-channel relationships, while triplet attention handles cross-dimensional interactions through parallel branches. Spatial attention focuses on spatially informative regions, and the final residual connection ensures stable gradient propagation. This hierarchical design enables feature enhancement across channel, spatial, and cross-dimensional aspects.

3.2.2. Identity-YOLO: Minimal Intervention Baseline

Identity-YOLO serves as a controlled baseline for evaluating Layer 10 enhancement mechanisms. To ensure fair comparison across all experimental variants, this study adopts a unified architectural framework where all models incorporate a processing module at Layer 10, positioned between the backbone (Layers 0–9) and neck (Layers 11–22). This design differs from the original YOLOv11n architecture, which directly connects the backbone to the neck without an intermediate processing layer.

As detailed in Figure 1b, the Identity-YOLO architecture maintains the standard YOLOv11n backbone and neck structure while introducing an IdentityNeck module at Layer 10. The IdentityNeck implements a pure identity mapping function (return x) that performs no computational operations on the input features of dimension [256, 20, 20]. This pass-through mechanism preserves the original feature representations without transformation, effectively replicating the information flow of standard YOLOv11n while maintaining architectural consistency with the other experimental variants.

The computational characteristics of Identity-YOLO establish it as an appropriate minimal intervention baseline. The IdentityNeck module contains zero trainable parameters and introduces negligible computational overhead (functionally equivalent to a direct connection), ensuring that any performance differences observed between Identity-YOLO and enhanced variants can be attributed solely to the enhancement mechanisms rather than to architectural inconsistencies or parameter budget differences. This controlled experimental design enables direct assessment of enhancement effectiveness while maintaining full compatibility with the YOLOv11n feature extraction and prediction pipeline.

The minimal intervention approach provides a reference point that isolates the contribution of feature enhancement processing at the backbone–neck transition. By eliminating enhancement operations while preserving the overall architectural framework, Identity-YOLO demonstrates the performance achievable through the standard YOLOv11n feature hierarchy when enhancement mechanisms are absent. This baseline enables quantitative evaluation of whether increased architectural complexity at Layer 10 yields proportional performance improvements across different dataset characteristics.

3.2.3. FPN-YOLO: Feature Pyramid Enhancement

FPN-YOLO implements a lightweight FPN enhancement mechanism at Layer 10, designed to improve multi-scale feature representation through efficient feature reduction and fusion operations. Feature pyramid architectures have demonstrated effectiveness for multi-scale object detection by building feature hierarchies with lateral connections [17]. As detailed in Figure 1b, the FPNDenseLite module replaces the standard transition layer with a streamlined pyramid processing pipeline that maintains computational efficiency while enhancing feature expressiveness. The architecture preserves the original YOLOv11n backbone and neck structure while introducing targeted feature pyramid operations at the critical backbone–neck interface.

The FPNDenseLite approach implements structured convolution-based processing through a three-stage pipeline: channel reduction, feature expansion with spatial context modeling, and multi-stream fusion. This design emphasizes computational efficiency by utilizing standard convolution operations and nearest-neighbor interpolation, making it suitable for resource-constrained scenarios where moderate feature enhancement is required without the complexity of attention-based mechanisms.

3.2.4. C2F-YOLO: Cross-Stage Partial Enhancement

C2F-YOLO incorporates a CSP enhancement mechanism at Layer 10 through the C2fNeck module, which implements efficient bottleneck processing with gradient flow optimization. The CSP design principle enables efficient gradient propagation while reducing computational redundancy [48]. As detailed in Figure 1b, the C2fNeck architecture employs a three-stage processing pipeline consisting of channel reduction, CSP processing, and channel expansion operations. The C2fNeck module implements CSP processing with $n = 2$ blocks and expansion ratio of 0.5, creating a bottleneck structure that reduces computational complexity while preserving essential feature information.

The final expansion stage employs a 1×1 convolution operation that transforms the processed 128-channel features back to the target output dimensionality of [256, 20, 20]. This channel expansion serves as a learned projection mechanism that integrates the CSP features into a unified representation suitable for subsequent neck processing. The C2fNeck design achieves feature enhancement through efficient bottleneck processing while maintaining computational tractability compared to more complex attention mechanisms.

3.2.5. Comparative Analysis and Architectural Trade-Offs

The four feature enhancement variants presented in Table 2 demonstrate distinct architectural philosophies for backbone–neck transition processing, each addressing different aspects of feature representation and computational complexity. The comparative analysis reveals fundamental trade-offs between feature preservation, enhancement sophistication, and processing demands based on their implementation characteristics.

Table 2. Comparative analysis of Layer 10 feature enhancement variants.

Model	Enhancement Type	Key Operations	Strengths	Limitations
HFE-YOLO	Multi-attention	CBAM (r = 16) + Triplet (k = 7) + Spatial + Residual	Multi-dimensional attention, Sequential processing	Higher computational cost
Identity-YOLO	Pass-through	Identity mapping	Minimal overhead, Direct gradient flow	No feature enhancement
FPN-YOLO	Lightweight FPN	Reduce (128)→Expand (256)→ Interp + Concat (384) →Fuse(256)	Structured processing, Multi-stream fusion	Limited attention mechanisms
C2F-YOLO	Cross-stage partial	Reduce (128)→C2f (n = 2, exp = 0.5)→Expand (256)	Bottleneck design, Structured processing	Moderate complexity

3.3. Datasets and Distribution Characteristics

This study evaluates the proposed models using two publicly available datasets from Roboflow Universe across complementary construction monitoring scenarios with fundamentally different class distribution characteristics. The Construction Project Monitoring dataset [49] addresses equipment detection tasks, while the MobileScaffoldingCheck dataset [50] focuses on safety compliance monitoring. These datasets are referred to as ConstructSight (CS) and SafeGuard (SG), respectively, throughout this study. Both datasets are publicly accessible with complete annotations: the CS dataset is available at https://universe.roboflow.com/ai-in-civil/02.-construction-project-monitoring (accessed on 20 June 2025) and the SG dataset at https://universe.roboflow.com/combinedmobilescaffolding/mobilescaffoldingcheck (accessed on 20 June 2025).

The selection of these datasets was driven by their contrasting distribution characteristics. The CS dataset exhibits pronounced class imbalance (manpower: 60.4%, diesel generator: 1.8%) characteristic of realistic construction monitoring scenarios. Construction monitoring research demonstrates persistent performance variations across equipment categories despite high overall accuracy, attributed to class-specific detection challenges and limited training representation in underrepresented categories [1,29,30]. These distribution characteristics allow evaluation of enhancement mechanisms under the studied conditions.

An overview of both datasets is provided to establish their characteristics. Table 3 provides an overview of both datasets, including the number of classes, images, instances, and object types. The CS dataset consists of 2195 images containing 7980 instances across eight construction equipment classes, while the SG dataset comprises 12,645 images with 26,601 instances across five safety-related classes.

Both datasets were preprocessed using standard object detection pipelines. For the CS dataset, images were resized to 640 × 640 pixels using stretch resizing with auto-orientation applied. Data augmentation included brightness adjustment (±15%) and blur (up to 0.5 px), generating two augmented versions per training example. For the SG dataset, images were resized to 640 × 640 pixels using fit resizing with black edges to preserve aspect ratios, with auto-orientation applied. More extensive augmentation

was applied including horizontal flipping, hue adjustment (±25°), saturation adjustment (±25%), brightness adjustment (±25%), and exposure adjustment (±25%), generating three augmented versions per training example. The augmentation strategies were selected to match dataset characteristics, with CS requiring moderate augmentation for equipment detection and SG employing extensive augmentation to improve generalization across diverse safety violation scenarios.

Table 3. Overview of datasets for construction equipment detection and safety compliance.

Task	Dataset	# Classes	# Images	# Instances	Object Classes
Equipment Detection	ConstructSight (CS)	8	2195	7980	Backhoe, concrete mixer, diesel generator, dump truck, excavator, manpower, tower hoist lift, tractor
Safety Compliance	SafeGuard (SG)	5	12,645	26,601	Missing guardrail, mobile scaffold (no outrigger), mobile scaffold (outrigger), worker with helmet, worker without helmet

Detailed statistics for the CS dataset are presented in Table 4. The train/val/test split ratio of 87.9/6.7/5.4% ensures sufficient training samples while maintaining representative validation and test sets. The average instances per image is 3.64 (calculated as total instances divided by total images). While the test set size (118 images) is smaller than conventional 20–30% allocations, it provides statistically adequate evaluation for object detection given the 455 total instances distributed across eight classes (7–291 instances per class), enabling reliable mAP calculation across precision-recall curves. This split configuration reflects the dataset's original design [49] and follows established practices in domain-specific object detection research where dataset availability constraints necessitate maximizing training data while maintaining evaluation validity [51].

Table 4. Dataset statistics for CS dataset.

Class	Train	Val	Test	Total	%
Backhoe	502	20	34	556	7.0
Concrete Mixer	356	34	21	411	5.2
Diesel Generator	124	10	7	141	1.8
Dump Truck	592	51	26	669	8.4
Excavator	526	52	27	605	7.6
Manpower	4108	417	291	4816	60.4
Tower Hoist Lift	206	25	12	243	3.0
Tractor	470	32	37	539	6.8
Total Instances	6884	641	455	7980	100.0
Total Images	1930	147	118	2195	—
Split Ratio (%)	87.9	6.7	5.4	100.0	—
Avg Inst./Image	3.57	4.36	3.86	3.64	—

The SG dataset statistics are shown in Table 5. Unlike the CS dataset, the SG dataset demonstrates a more balanced class distribution, with class frequencies ranging from 15.2% (missing guardrail) to 25.9% (worker with helmet). The dataset employs a train/val/test split ratio of 92.4/3.8/3.8%, with a higher training proportion to provide sufficient samples for learning diverse safety violation patterns. The average of 2.10 instances per image is

lower than the CS dataset, reflecting the nature of safety violation detection where specific violations may occur less frequently within individual frames.

Table 5. Dataset statistics for SG dataset.

Class	Train	Val	Test	Total	%
Missing Guardrail	3723	146	169	4038	15.2
Mobile Scaffold (No Outrigger)	3897	176	114	4187	15.7
Mobile Scaffold (Outrigger)	4435	186	173	4794	18.0
Worker with Helmet	6124	328	442	6894	25.9
Worker without Helmet	5983	224	481	6688	25.1
Total Instances	24,162	1060	1379	26,601	100.0
Total Images	11,679	483	483	12,645	—
Split Ratio (%)	92.4	3.8	3.8	100.0	—
Avg Inst./Image	2.07	2.19	2.86	2.10	—

The SG dataset employs a 92.4/3.8/3.8% split ratio, allocating a higher proportion to training compared to the CS dataset to accommodate the larger dataset scale and diverse safety violation patterns. The test set (483 images, 1379 instances) provides robust evaluation capability with substantially higher instance counts than the CS dataset, ensuring adequate per-class representation (114–481 instances per class) for reliable performance assessment. The smaller test percentage (3.8% vs. CS 5.4%) is offset by the larger absolute instance count, maintaining evaluation validity through instance-level adequacy consistent with domain-specific object detection practices [51]. This dual-dataset design with complementary split configurations enables architectural comparison across different data scales while maintaining statistically adequate evaluation for both imbalanced (CS) and balanced (SG) distribution scenarios.

3.4. Experimental Environment

All experiments were conducted on a workstation equipped with an Intel Xeon W-2265 CPU, NVIDIA RTX 3090 GPU with 24 GB memory, and 64 GB system RAM, running Windows 11 operating system. The implementation utilized Python 3.9.7 with PyTorch 2.2.2 deep learning framework and CUDA 11.8 for GPU acceleration. Additional dependencies include NumPy 1.24.3, OpenCV 4.8.0, and Matplotlib 3.7.1 for data processing and visualization.

All models were trained for 200 epochs with a batch size of 16 and input resolution of 640 $\times$ 640 pixels in RGB format. The initial learning rate was set to 0.01 with automatic optimizer selection between AdamW and SGD based on model characteristics.

3.5. Performance Metrics

The performance of the proposed models is evaluated using standard object detection metrics that assess both localization accuracy and classification performance. These metrics follow the COCO evaluation protocol [52] and are widely adopted in the object detection community.

3.5.1. Intersection over Union (IoU)

IoU measures the overlap between predicted and ground truth bounding boxes (Equation (1)):

$$\text{IoU} = \frac{\text{Area}\left(B_{pred} \cap B_{gt}\right)}{\text{Area}\left(B_{pred} \cup B_{gt}\right)} \tag{1}$$

where B_{pred} represents the predicted bounding box and B_{gt} represents the ground truth bounding box.

3.5.2. Precision and Recall

Precision quantifies the proportion of correct detections among all predictions (Equation (2)):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall measures the proportion of ground truth objects that are successfully detected (Equation (3)):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

where TP (True Positives) represents correctly detected objects, FP (False Positives) represents incorrect detections, and FN (False Negatives) represents missed objects. A detection is considered a true positive when IoU $\geq$ threshold (typically 0.5).

3.5.3. Average Precision (AP) and Mean Average Precision (mAP)

Average precision (AP) summarizes the precision-recall curve for a single class (Equation (4)):

$$\text{AP} = \int_0^1 p(r)dr \tag{4}$$

where $p(r)$ represents precision as a function of recall. In practice, AP is computed as the area under the precision-recall curve using the interpolated precision at each recall level.

Mean average precision (mAP) extends AP to multiple classes by averaging across all object categories (Equation (5)):

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \tag{5}$$

where N is the number of classes and AP_i is the average precision for class i.

3.5.4. mAP@50 and mAP@50–95

Two primary mAP variants are reported in this study. mAP@50 represents mean average precision at IoU threshold of 0.5, providing standard detection accuracy evaluation. mAP@50–95 represents mean average precision averaged over IoU thresholds from 0.5 to 0.95 with step size 0.05, providing a more stringent evaluation of localization quality. The mAP@50–95 metric is calculated as Equation (6):

$$\text{mAP@50–95} = \frac{1}{10}\sum_{t=0.5}^{0.95} \text{mAP@}t \tag{6}$$

where t represents IoU thresholds {0.5, 0.55, 0.60, . . ., 0.95}.

3.5.5. Loss Functions

The model training employs a composite loss function combining three components (Equation (7)):

$$L_{total} = \lambda_{box} L_{box} + \lambda_{cls} L_{cls} + \lambda_{dfl} L_{dfl} \quad (7)$$

where L_{box} is the bounding box regression loss based on the Complete IoU (CIoU) metric, which accounts for overlap area, center distance, and aspect ratio, L_{cls} is the classification loss using binary cross-entropy for multi-class prediction, L_{dfl} is the distribution focal loss for box refinement through distribution modeling, λ_{box} = 7.5, λ_{cls} = 0.5, and λ_{dfl} = 1.5.

These loss weights follow the default YOLOv11 configuration. The relatively higher box loss weight emphasizes accurate localization, appropriate for construction site monitoring applications requiring accurate object localization.

4. Experimental Results

4.1. Performance on CS Dataset

The training and validation curves for all models on the CS dataset are presented in Figure 2. All models demonstrate consistent convergence, with training losses decreasing steadily across epochs. Among the loss components, box loss and DFL loss show faster convergence compared to classification loss.

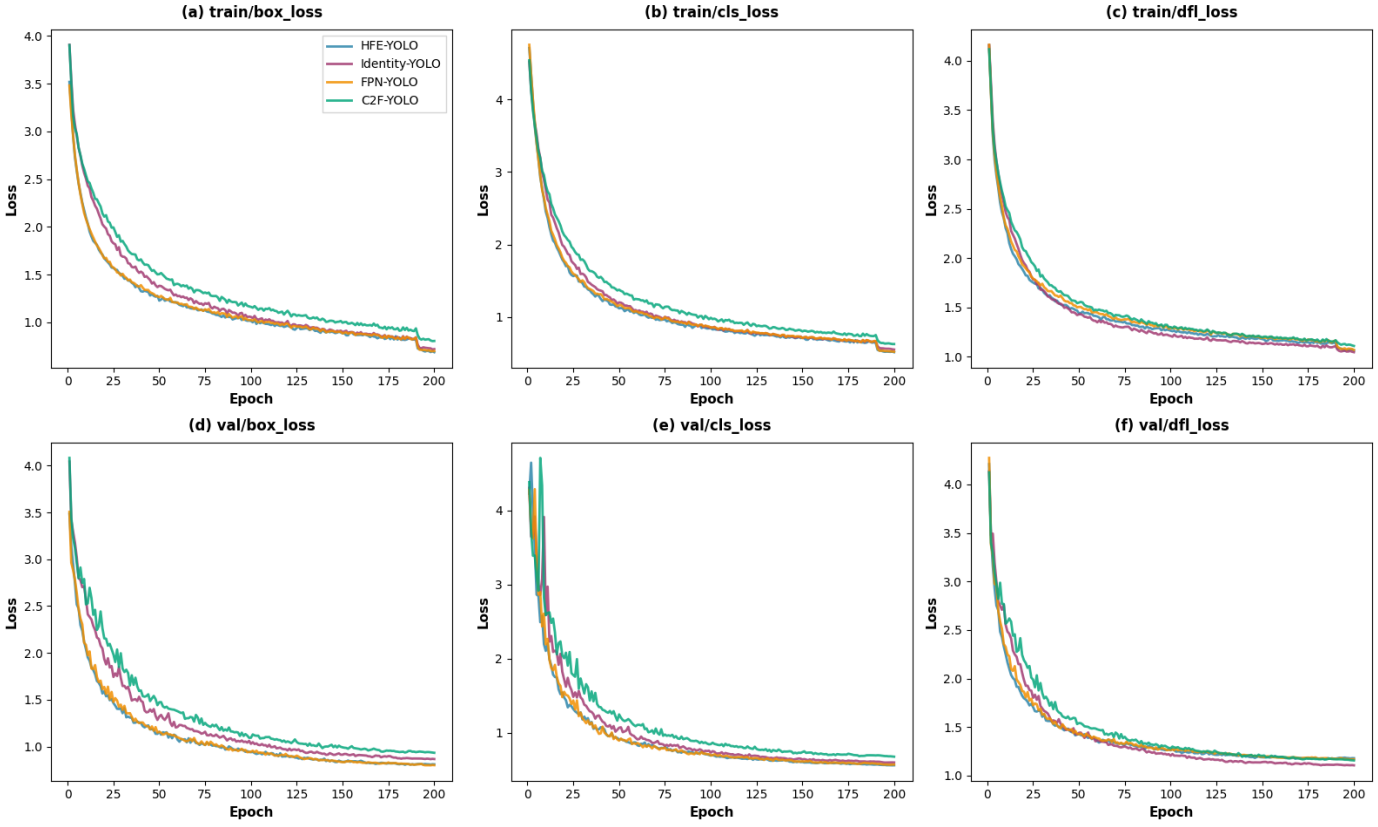


Figure 2. Training and validation loss comparison on CS dataset.

Performance metrics across training epochs are shown in Figure 3. All models achieve high precision (>0.91) and recall (>0.89) values, showing good detection performance. The mAP@50 metric converges to values exceeding 0.92 for all models.

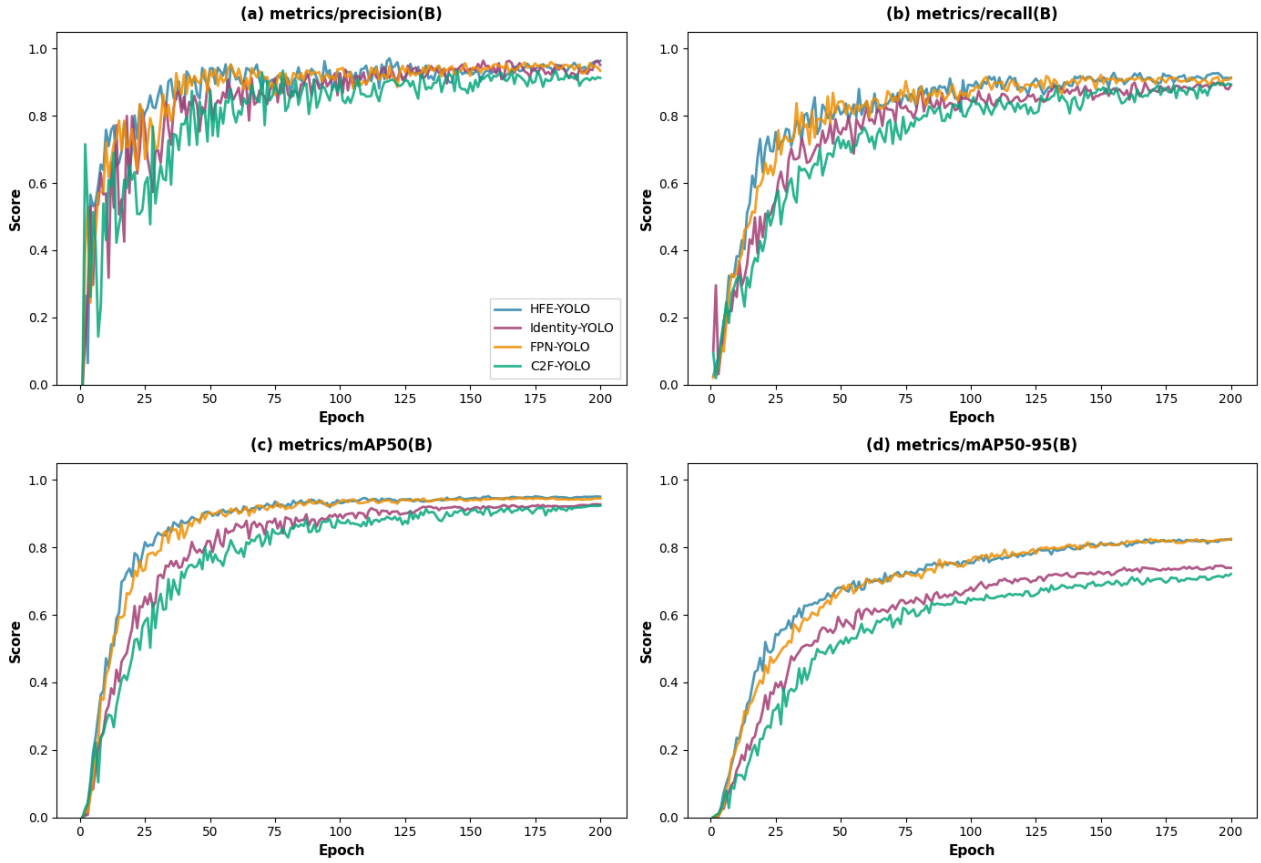


Figure 3. Performance metrics comparison on CS dataset.

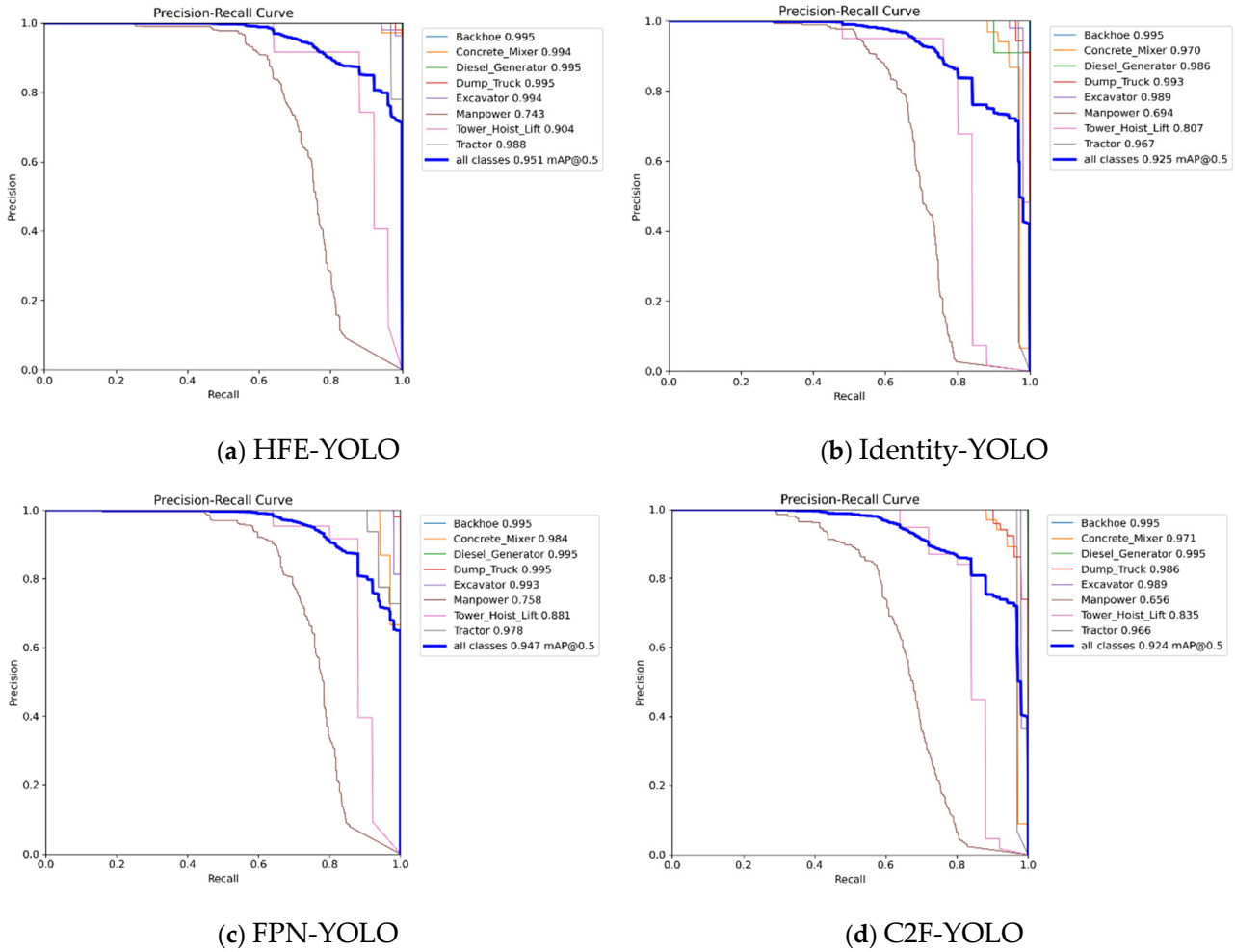
Table 6 presents the performance summary for all models on the CS dataset. HFE-YOLO achieves the best overall performance, ranking first in all four metrics: precision (0.959), recall (0.912), mAP@50 (0.950), and mAP@50–95 (0.826), closely followed by FPN-YOLO (0.824 mAP@50–95). Results show improved performance with multi-attention mechanisms [38,47] and multi-scale feature fusion through structured pyramid processing [17] for the imbalanced CS dataset. Identity-YOLO and C2F-YOLO show lower performance, with mAP@50–95 of 0.744 and 0.721, respectively.

Table 6. Model performance summary on CS dataset.

Model	Precision	Recall	mAP@50	mAP@50–95
HFE-YOLO	0.959	0.912	0.950	0.826
FPN-YOLO	0.950	0.907	0.948	0.824
Identity-YOLO	0.939	0.890	0.925	0.744
C2F-YOLO	0.913	0.897	0.924	0.721

Figure 4 presents the precision-recall curves for per-class performance evaluation on the CS dataset. The curves reveal significant performance variation across different object categories, corresponding to the impact of class imbalance in the training data. As shown in Figure 4a, HFE-YOLO achieves the highest overall mAP@0.5 of 0.951, with most equipment classes demonstrating AP values exceeding 0.99. However, the manpower class exhibits notably lower performance with AP of 0.743, despite representing 60.4% of the training instances. This performance gap persists across all model variants, as demonstrated in

Figure 4b–d, where Identity-YOLO, FPN-YOLO, and C2F-YOLO achieve manpower AP values of 0.694, 0.758, and 0.656, respectively.



(**a**) HFE-YOLO (**b**) Identity-YOLO

(**c**) FPN-YOLO (**d**) C2F-YOLO

Figure 4. Per-class precision-recall curves on CS dataset.

The equipment classes show consistently high performance across all models, with backhoe, diesel generator, dump truck, and excavator achieving AP values above 0.98 in most cases. Figure 4c shows that FPN-YOLO achieves particularly strong performance on equipment classes, with several classes reaching AP of 0.995. The tower hoist lift class presents moderate detection difficulty, with AP values ranging from 0.807 (Identity-YOLO in Figure 4b) to 0.904 (HFE-YOLO in Figure 4a), indicating intermediate complexity in visual features and pose variations. The precision-recall curves demonstrate that all models maintain high precision across most recall levels for equipment classes, while the manpower class shows steeper precision degradation as recall increases, suggesting challenges in detecting all worker instances while maintaining classification confidence.

Figure 5 presents qualitative detection results on a representative evaluation image from the CS dataset, comparing the output of all four models. The visualization demonstrates that all architectures successfully detect major equipment such as excavators and dump trucks with high confidence scores (>0.9), while showing varying performance on smaller objects and personnel detection.



(**a**) HFE-YOLO (**b**) Identity-YOLO

(**c**) FPN-YOLO (**d**) C2F-YOLO

Figure 5. Representative detection results on CS dataset. Detection results from (**a**) HFE-YOLO, (**b**) Identity-YOLO, (**c**) FPN-YOLO, and (**d**) C2F-YOLO on the same image. Colored bounding boxes indicate detected objects with class labels and confidence scores.

4.2. Performance on SG Dataset

Figure 6 presents the training and validation loss curves for the SG dataset. The convergence patterns differ notably from the CS dataset, with all models showing more stable validation losses throughout training. This behavior coincides with the more balanced class distribution in the SG dataset. The loss components converge at similar rates, suggesting that both localization and classification tasks are equally challenging for safety violation detection.

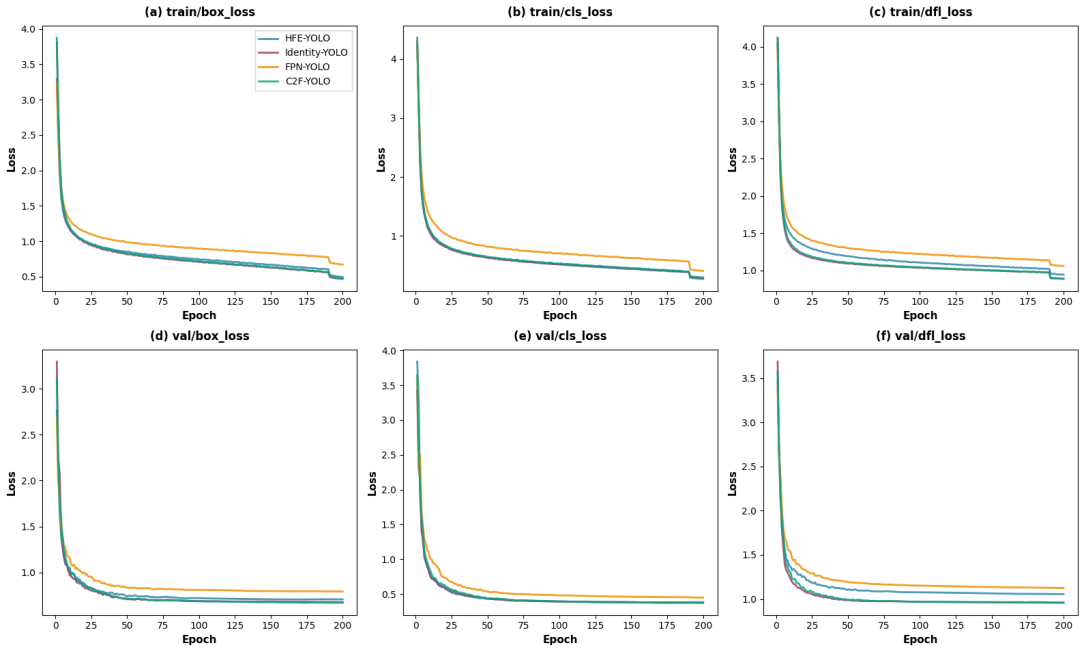


Figure 6. Training and validation loss comparison on SG dataset.

The performance metrics evolution is illustrated in Figure 7. Models achieve higher precision and recall on the SG dataset compared to the CS dataset, with precision exceeding 0.94 and most models surpassing 0.94 for recall. The mAP@50 values consistently surpass 0.96 for HFE-YOLO and FPN-YOLO during training, indicating strong detection performance. However, the mAP@50–95 values (0.78–0.80) are similar to those observed on the CS dataset, suggesting consistent challenges in tight bounding box predictions across both domains.

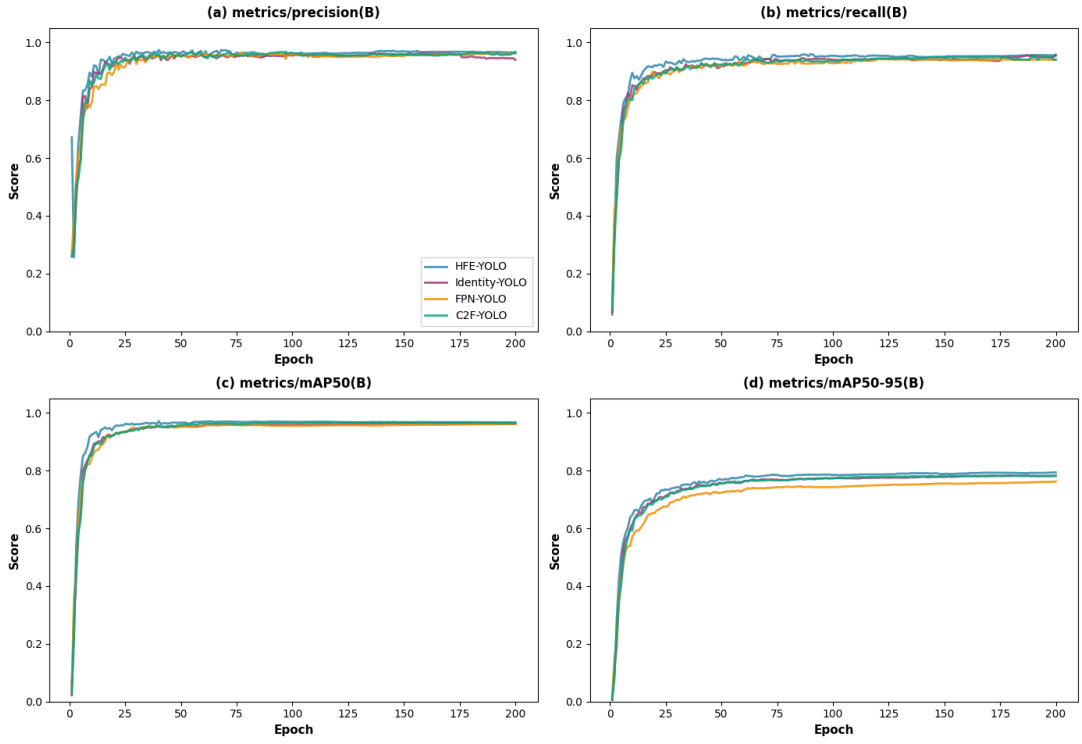


Figure 7. Performance metrics comparison on SG dataset.

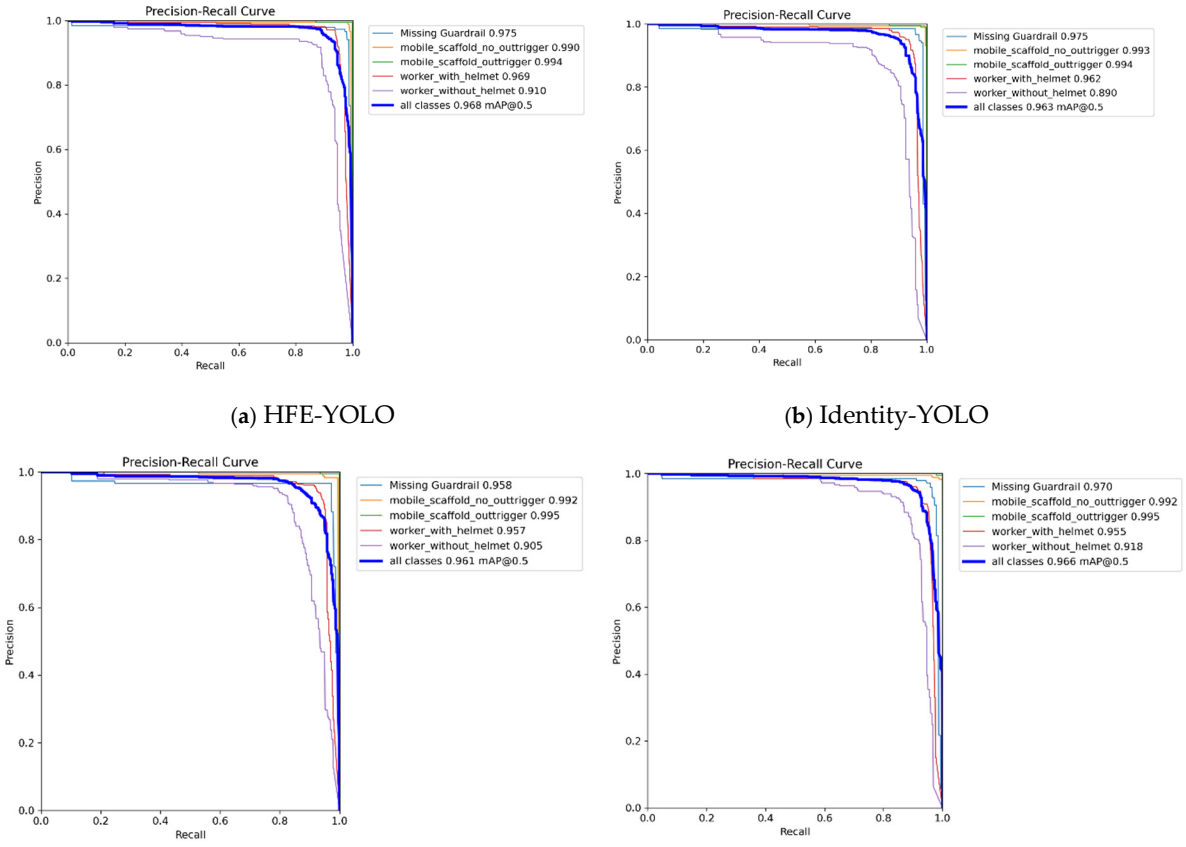
Table 7 presents the performance summary for the SG dataset. HFE-YOLO achieves superior performance across all metrics, attaining the highest precision (0.964), recall (0.956), mAP@50 (0.968), and mAP@50–95 (0.794). C2F-YOLO and Identity-YOLO demonstrate competitive performance with mAP@50–95 of 0.782 and 0.781, respectively.

Table 7. Model performance summary on SG dataset.

Model	Precision	Recall	mAP@50	mAP@50–95
HFE-YOLO	0.964	0.956	0.968	0.794
FPN-YOLO	0.962	0.940	0.961	0.761
C2F-YOLO	0.958	0.948	0.966	0.782
Identity-YOLO	0.949	0.948	0.963	0.781

Notably, FPN-YOLO achieves lower performance (0.761 mAP@50–95) on the balanced SG dataset compared to the imbalanced CS dataset, indicating sensitivity to class distribution characteristics. All models demonstrate notably higher precision and recall values compared to the CS dataset, with precision exceeding 0.94 across all architectures. This improvement accompanies the more balanced class distribution (15.2–25.9%) and focused safety violation categories in the SG dataset.

Figure 8 presents the precision-recall curves for all models on the SG dataset, illustrating per-class detection performance and the overall mAP@50 metric. HFE-YOLO achieves mAP@50 of 0.968, Identity-YOLO achieves 0.963, FPN-YOLO achieves 0.961, and C2F-YOLO achieves 0.966, all showing consistent performance across safety violation categories.



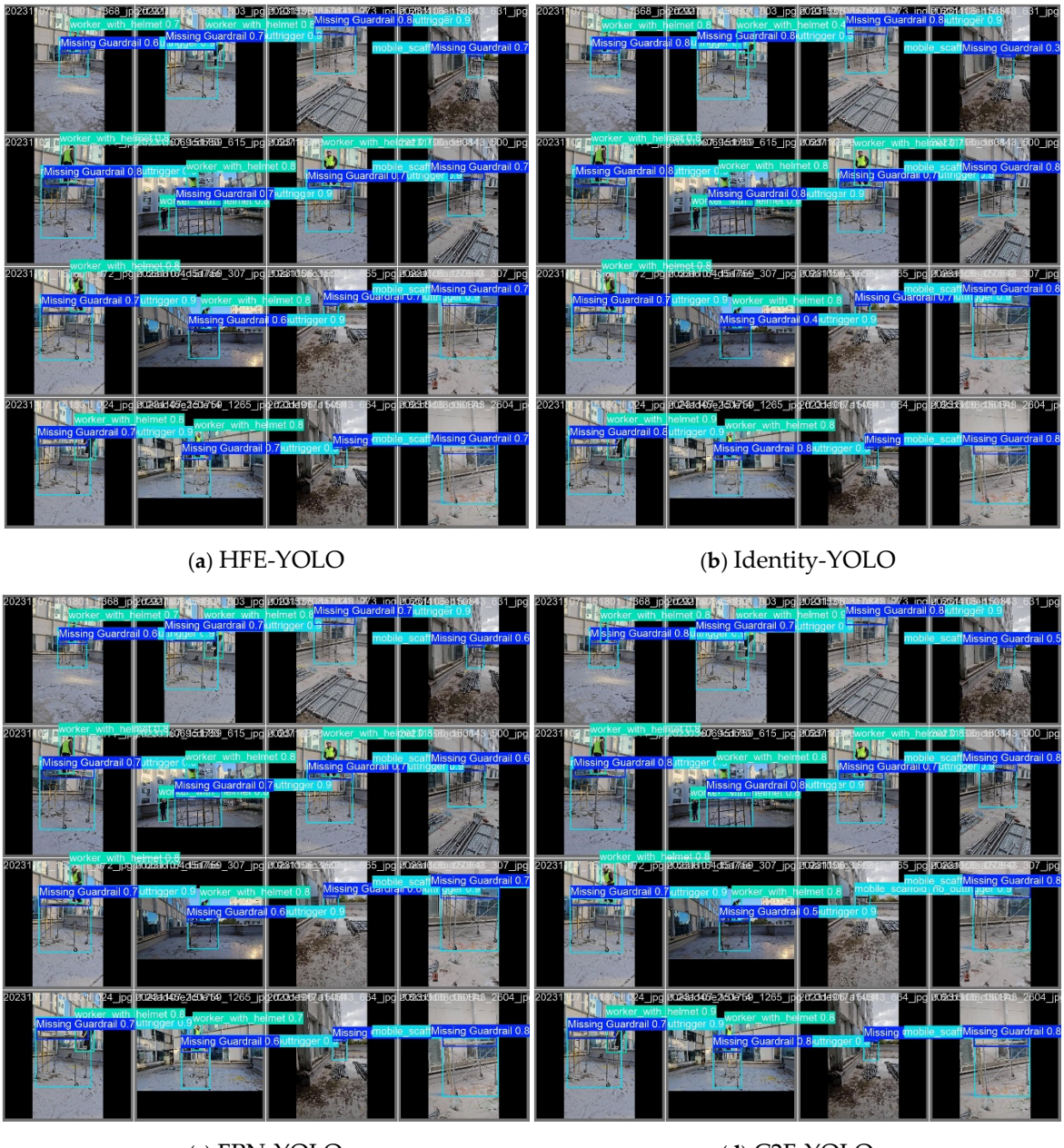
(**a**) HFE-YOLO (**b**) Identity-YOLO

(**c**) FPN-YOLO (**d**) C2F-YOLO

Figure 8. Per-class precision-recall curves on SG dataset.

On evaluation, all models achieve high performance, with mAP@50–95 values ranging from 0.761 to 0.794—narrower than the 0.721–0.826 range observed on the CS dataset. This compression of performance differences indicates that enhancement mechanisms show greater benefit on the imbalanced dataset compared to the balanced dataset.

Representative detection examples from the SG dataset evaluation samples are shown in Figure 9. The models successfully detect various safety violations including missing guardrails, scaffold configurations, and worker helmet compliance. Multiple detections per image reflect the average of 2.10 instances per image in the SG dataset.



(**a**) HFE-YOLO (**b**) Identity-YOLO

(**c**) FPN-YOLO (**d**) C2F-YOLO

Figure 9. Representative detection results on SG dataset. Examples show model predictions with confidence scores for safety violation detection including missing guardrail (blue), mobile scaffold configurations (cyan), and worker helmet compliance (turquoise).

4.3. Enhancement Mechanism Comparison at Layer 10

This section presents a controlled comparison of four enhancement strategies at the backbone–neck transition point (Layer 10), evaluated against Identity-YOLO as a minimal

intervention baseline. All architectures share identical backbone (Layers 0–9), neck (Layers 11–22), and head (Layer 23) structures, differing only in Layer 10 processing. This controlled design enables direct assessment of enhancement mechanism effectiveness while maintaining consistent feature extraction and prediction pipelines.

4.3.1. CS Enhancement Comparison

Table 8 presents the enhancement mechanism comparison on the CS dataset using performance metrics, quantifying the performance contribution of each strategy relative to the Identity-YOLO baseline. The results reveal substantial performance variations depending on the enhancement approach, with differences ranging from −2.3 percentage points to +8.2 percentage points in mAP@50–95.

Table 8. Enhancement mechanism comparison on CS dataset.

Model	Enhancement Type	mAP@50–95	Δ vs. Identity	Precision	Recall	mAP@50
Identity-YOLO	Minimal (pass-through)	0.744	reference	0.939	0.890	0.925
C2F-YOLO	Cross-stage partial	0.721	−0.023	0.913	0.897	0.924
FPN-YOLO	Feature pyramid	0.824	+0.080	0.950	0.907	0.948
HFE-YOLO	Multi-attention	0.826	+0.082	0.959	0.912	0.950

HFE-YOLO achieves the highest mAP@50–95 of 0.826 (+8.2 percentage points), with FPN-YOLO achieving the second-highest performance (0.824, +8.0 percentage points). This substantial gain shows improved performance with multi-attention mechanisms [38,47] and multi-scale feature fusion through structured pyramid processing [17] for the imbalanced CS dataset.

Notably, C2F-YOLO exhibits degraded performance compared to Identity-YOLO (−2.3 percentage points), suggesting that CSP bottleneck processing may be suboptimal for highly imbalanced datasets where the majority class (manpower, 60.4%) dominates the feature space.

4.3.2. SG Enhancement Comparison

Table 9 presents the enhancement mechanism comparison on the SG dataset using performance metrics, revealing markedly different effectiveness patterns compared to the CS dataset.

Table 9. Enhancement mechanism comparison on SG dataset.

Model	Enhancement Type	mAP@50–95	Δ vs. Identity	Precision	Recall	mAP@50
Identity-YOLO	Minimal (pass-through)	0.781	reference	0.949	0.948	0.963
C2F-YOLO	Cross-stage partial	0.782	+0.001	0.958	0.948	0.966
FPN-YOLO	Feature pyramid	0.761	−0.020	0.962	0.940	0.961
HFE-YOLO	Multi-attention	0.794	+0.013	0.964	0.956	0.968

These compressed performance differences on the balanced SG dataset (1.3 percentage point range from Identity to HFE) contrast sharply with the CS dataset results (8.2 percentage point gap for HFE-YOLO), indicating that enhancement mechanism selection has substantially reduced importance on balanced datasets.

4.3.3. Enhancement Mechanism Analysis

The comparative evaluation reveals key findings regarding enhancement mechanism effectiveness across different dataset characteristics. A notable finding concerns baseline performance on balanced datasets. Identity-YOLO achieves competitive performance on the balanced SG dataset (0.782 mAP@50–95, 0.963 precision, 0.958 recall), approaching sophisticated enhancement variants within 1.7% relative difference. This result suggests that the minimal intervention baseline achieved competitive performance on the balanced SG dataset. The strong baseline performance suggests that computational resources allocated to complex enhancement mechanisms may yield diminishing returns on well-balanced datasets, where simpler architectures can achieve comparable detection quality with lower computational overhead. In contrast, CSP-based enhancement shows limited effectiveness across both datasets. C2F-YOLO consistently underperforms or marginally exceeds the Identity-YOLO baseline across both datasets (−2.3 percentage points on CS, +0.1 on SG), despite implementing established CSP design principles [48]. The bottleneck structure with $n = 2$ blocks and 0.5 expansion ratio appears insufficient for effective feature enhancement at the backbone–neck transition point.

5. Discussion

5.1. Enhancement Mechanism Effectiveness

The systematic comparison provides insights into backbone–neck transition enhancement based on dataset distribution characteristics, extending prior work on attention mechanisms in object detection [10,25].

The empirical findings provide direct evidence addressing RQ1 (Enhancement Mechanism Effectiveness) and RQ2 (Distribution-Dependent Performance), demonstrating that enhancement effectiveness depends fundamentally on data distribution characteristics rather than representing universal architectural improvements.

These findings reveal several key patterns. First, enhancement mechanisms demonstrate varying effectiveness across dataset characteristics. On the imbalanced CS dataset, feature pyramid and multi-attention enhancements provide substantial improvements (+8.0 and +8.2 percentage points), while on the balanced SG dataset, the same mechanisms yield modest gains (+1.3 percentage points for HFE-YOLO). These findings align with class imbalance literature showing that sophisticated architectures provide greater benefits when class distributions are skewed [53–55].

The implications for architecture selection are clear. For the evaluated scenarios with severe class imbalance, sophisticated enhancement mechanisms demonstrated substantial improvements on the tested datasets. The effectiveness of multi-scale processing through feature pyramid architectures corroborates findings from Lin et al. [17] regarding FPN effectiveness, while multi-dimensional attention results support recent work on attention mechanisms for object detection [38,47].

The superior performance of structured feature transformation approaches compared to bottleneck-based processing on these specific datasets indicates different effectiveness patterns for the evaluated architectures. At the backbone–neck transition point where features must support multi-scale detection across pyramid levels, preservation and refinement operations showed better performance than compression strategies in this study.

The C2F-YOLO performance pattern warrants specific analysis. On the imbalanced CS dataset, C2F-YOLO achieves 72.1% mAP@50–95, underperforming the Identity-YOLO baseline (74.4%) by 2.3 percentage points despite implementing established CSP design principles [48]. This degradation likely stems from the bottleneck architecture's feature compression characteristics. The C2fNeck module employs channel reduction to 128 dimensions (50% compression ratio) followed by CSP processing with $n = 2$ blocks, creating

an information bottleneck at the critical backbone–neck transition point. For severely imbalanced datasets where the majority class (manpower: 60.4%) dominates the feature space, this compression may disproportionately suppress minority class representations essential for detecting underrepresented equipment categories. The bottleneck design, while effective for computational efficiency in balanced scenarios [48], appears suboptimal when class imbalance introduces asymmetric feature importance distributions. In contrast, feature preservation and refinement approaches—implemented by Identity-YOLO (zero compression) and enhancement mechanisms like FPN-YOLO and HFE-YOLO (expansion-based processing)—maintain richer feature representations that better accommodate imbalanced class distributions. This pattern suggests that architectural choices at critical transition points should account for dataset distribution characteristics, with compression-based approaches potentially requiring adaptive mechanisms to preserve minority class features in imbalanced scenarios.

The C2F-YOLO negative result provides counter-evidence to established CSP effectiveness [48], contributing valuable constraints for future enhancement mechanism selection in imbalanced detection scenarios.

Conversely, for scenarios with balanced class distributions where all categories receive adequate training representation, the minimal intervention baseline achieves a performance approaching sophisticated enhancement variants. This compression of architectural differences supports the hypothesis that class distribution characteristics significantly influence architectural choice effectiveness.

The consistent ranking preservation across datasets despite varying absolute improvements indicates that while enhancement magnitude varies, relative architectural effectiveness remains stable. Based on the tested conditions, feature pyramid mechanisms provide consistent performance across the evaluated datasets.

5.2. Class-Specific Performance

The per-class performance analysis reveals patterns that warrant further investigation, building on class imbalance literature [53,54]. These patterns address RQ3 (Integrated Monitoring Capability) by revealing fundamental differences in detection difficulty across equipment and safety monitoring tasks within the unified architectural framework.

The analysis reveals distinct patterns across different distribution scenarios. On the imbalanced dataset, the majority class exhibits substantially lower detection performance compared to minority equipment classes despite dominating training instances. This observation is consistent with findings in computer vision showing that class frequency alone does not guarantee detection performance [17]. Worker detection faces challenges including posture variations, movement variations, occlusion scenarios, and multi-scale detection difficulties in complex construction environments [56–58].

In contrast, equipment classes with substantially lower training frequencies achieve consistently high detection performance. These rigid mechanical objects maintain stable geometric structures, distinctive color patterns, and characteristic shapes that remain consistent across instances. This pattern aligns with object detection literature showing that visual consistency often outweighs sample quantity [59].

In contrast, on the balanced dataset, the worker helmet compliance class presents the most challenging detection despite balanced representation. This finding is consistent with fine-grained object detection literature emphasizing the difficulty of detecting small, variable features [19]. Scaffold-related classes consistently achieve highest detection performance, confirming that geometric structures with clear boundaries provide robust visual cues as established in prior work [4].

This finding validates the unified enhancement mechanism approach for RQ3, demonstrating that structural safety monitoring does not inherently require specialized architectures distinct from equipment detection when target objects possess geometric consistency.

5.3. Comparative Benchmarking with Prior Literature

To contextualize the experimental findings, this subsection benchmarks the results against representative construction monitoring studies from Table 1, examining relative performance positioning, computational characteristics, and methodological trade-offs not addressed in the preceding analysis.

Performance positioning relative to specialized studies reveals competitive standing within established ranges. For equipment detection, the 95.0% mAP@50 achieved on CS dataset (eight classes, 2195 images) positions between Eum et al. [29] at 89.72% across five heavy equipment classes (21,772 images) and Shin et al. [1] at 98.93% across nine equipment classes (10,294 images). For safety compliance, the 96.8% mAP@50 on SG dataset (five classes, 12,645 images) approaches Abbas et al. [4] at 97.9% for scaffolding safety (five classes, 4868 images). These comparisons establish that the controlled architectural comparison achieves detection accuracy within the performance envelope of specialized single-task studies while addressing integrated equipment-safety monitoring.

Computational efficiency analysis reveals favorable characteristics for deployment scenarios. The YOLOv11n baseline provides 4.7 M parameters and 9.3 GFLOPs computational demands, substantially lower than architectures employed in comparable studies. Shin et al. [1] utilize YOLOv5l with higher parameter counts, while Liu et al. [26] implement multi-task YOLOv8 requiring segmentation and pose estimation beyond detection capabilities. The enhancement mechanisms evaluated introduce minimal overhead: HFE-YOLO's multi-attention processing operates on 256-channel features at 20×20 resolution, maintaining real-time processing capability while providing the documented 8.2 percentage point improvement on imbalanced data. Recent deployment studies report 92.41 FPS [60] and 123.6 FPS [61] for specialized applications, establishing computational feasibility benchmarks that YOLOv11n-based architectures can approach through efficient base architecture design.

Methodological trade-offs distinguish this investigation from prior work through controlled isolation emphasis. Unlike studies comparing complete architectural modifications [25,27] or different base models (YOLOv5l [1], YOLOv10 [29]), maintaining identical backbone–neck–head components across all variants enables direct enhancement mechanism attribution. Three specific trade-offs emerge: (1) architectural comparison priority over absolute performance maximization, yielding mechanistic insights at the cost of potential performance gains from specialized optimizations; (2) standard RGB input rather than depth sensing [4] or LiDAR integration, favoring deployment flexibility over specialized sensing capabilities; and (3) single enhancement position evaluation (Layer 10) rather than comprehensive modifications [10,19], providing controlled comparison at the expense of potential synergistic benefits from multi-point enhancement.

The dual-dataset evaluation addressing contrasting class distributions represents a methodological contribution absent from compared studies, which predominantly employ single datasets. This design quantifies distribution-dependent effectiveness (8.2 vs. 1.3 percentage points improvement magnitude), providing actionable guidance for architecture selection based on application-specific data characteristics—an insight unavailable from single-distribution evaluations predominant in the existing literature.

5.4. Broader Implications and Limitations

These findings have several broader implications for object detection research. The observed distribution-dependent effectiveness patterns suggest that architecture selection

should consider dataset characteristics rather than applying universal solutions, extending work by Liu et al. [26] on multi-task frameworks.

However, several limitations should be acknowledged. The findings are derived from two specific construction monitoring datasets representing distinct distribution characteristics. While these datasets provide controlled comparison conditions, generalization requires validation across broader construction monitoring scenarios. The study employs a minimal intervention baseline rather than unmodified YOLOv11n architecture, which may affect direct comparability with published benchmarks.

Additionally, the study focuses specifically on backbone–neck transition enhancement without exploring other architectural components, limiting the scope of architectural insights. Consistent performance plateaus at stricter IoU thresholds indicate persistent challenges in precise bounding box localization across all tested architectures.

Based on these limitations, several promising research directions emerge: (1) evaluation across diverse construction monitoring tasks including indoor renovation, night construction, and adverse weather scenarios; (2) investigation of alternative enhancement positions within YOLO architectures; (3) development of adaptive enhancement mechanisms that adjust based on detected class distribution characteristics; and (4) exploration of pose-aware detection frameworks for improved worker detection in construction environments.

6. Conclusions

This study systematically evaluated four feature enhancement mechanisms at the YOLOv11n backbone–neck transition using construction monitoring datasets with contrasting class distributions. The investigation provides empirical answers to three research questions:

RQ1 (Enhancement Mechanism Effectiveness): Feature enhancement mechanisms demonstrate varying effectiveness, with HFE-YOLO achieving superior performance (mAP@50: 95.0% on ConstructSight, 96.8% on SafeGuard) followed by FPN-YOLO (mAP@50: 94.8%, 96.1%), while CSP-based enhancement underperforms the baseline on imbalanced data (-2.3 percentage points). Enhancement magnitude depends on both mechanism type and data characteristics.

RQ2 (Distribution-Dependent Performance): Enhancement effectiveness shows strong dependency on dataset distribution, with substantial improvements on imbalanced data (+8.2 percentage points mAP@50–95 for HFE-YOLO on ConstructSight) compared to modest gains on balanced distributions (+1.3 percentage points on SafeGuard). This six-fold reduction in improvement magnitude indicates that architectural complexity should be calibrated to dataset characteristics rather than universally applied.

RQ3 (Integrated Monitoring Capability): The same architectural framework demonstrates effective performance across both equipment detection and safety compliance tasks (95.0–96.8% mAP@50), establishing the viability of unified enhancement mechanisms for diverse construction monitoring applications without task-specific architectural adaptations.

These findings contribute empirical evidence regarding enhancement mechanism selection based on dataset distribution characteristics and task requirements, offering insights for researchers working with class-imbalanced detection tasks and integrated construction monitoring systems.

Funding: This research budget was allocated by National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok (Project no. KMUTNB-FF-68-B-64.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the author.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table 1. Comprehensive literature review of construction site monitoring systems (2021–2025).

Study	Detection Target	Base Architecture	Key Technical Innovations	Classes	Dataset (Size)	mAP@0.5 (%)	Notable Contribution	Limitations
[8]	Construction machinery	Faster-RCNN-ResNet101	First construction-dataset (ACID); 4-algorithm benchmark	10 classes: Excavator, dump truck, compactor, dozer, grader, concrete mixer truck, wheel loader, backhoe loader, tower crane, mobile crane	10,000 images	89.2	First comprehensive machine dataset with systematic methodology	Limited 10 machine types; annotation quality issues
[28]	MOCS: Moving objects	Large-scale real sites	174 sites across multiple countries and project types	13 classes: Worker, tower crane, hanging hook, vehicle crane, roller, bulldozer, excavator, truck, loader, pump truck, concrete mixer truck, pile driver, other vehicle	41,668 images	51.04	Largest moving object dataset; 174-site diversity; systematic benchmark	Scaffold work mentioned but no safety monitoring class
[31]	SODA: 4M1E construction	Multi-device collection	UAV + handheld cameras; multiple construction phases; privacy protection	15 classes: Worker, helmet, vest, board, wood, rebar, brick, scaffold, handcart, cutter, electric box, hopper, hook, fence, slogan	19,846 images (10 sites: Guangzhou, Shenzhen, Dongguan)	81.47	Largest 4M1E dataset; per-pixel segmentation; publicly available	Scaffold lacks safety compliance specifications
[5]	Machinery swarm operations	Improved YOLOv4	K-means anchors; hybrid dilated convolution; Focal loss	4 classes: Loader, excavator, truck, person	Self-built swarm operations (11,250 instances)	97.03	First swarm operations focus; maintained 31.11 FPS	Occlusion limitations; dataset not public
[9]	Multi-scale site monitoring	YOLOv5s + SOD	Small object detection; edge computing deployment	16 classes: Excavator, payloader, forklift, dump truck, mixer truck, pump car, pile driver, truck, person, bicycle, car, motorcycle, bus, scissor lift, tower crane, aerial lift truck	COCO + AI Hub (299,655 images, 575,913 instances)	75.6	First edge computing implementation	Manual configuration; limited equipment-safety integration
[1]	Heavy equipment benchmark	YOLOv5l	K-means anchors; systematic augmentation	9 classes: Bulldozers, dump trucks, excavators, forklifts, loaders, MEWP, mobile cranes, Remicon, scrapers	Web-crawled + onsite (10,294 images)	98.93 (mAP@0.5:0.95: 90.26)	Highest reported accuracy (96.61–99.79% per-class AP)	Poor low-light performance; dataset not public

Table 1. *Cont.*

Study	Detection Target	Base Architecture	Key Technical Innovations	Classes	Dataset (Size)	mAP@0.5 (%)	Notable Contribution	Limitations
[25]	Occluded object detection	YOLOv7 + ECA	Efficient Channel Attention; SIoU loss; DIoU-NMS	6 classes: Workers, machinery, vehicles, excavators, raw materials, construction components	MOCS dataset (17,000 images)	85.4	Comprehensive occlusion handling mechanisms	Poor small object performance; no real-world validation
[26]	Multi-task safety monitoring	Multi-task YOLOv8	Unified detection + segmentation + pose + tracking	7 classes: Workers, excavators, cranes, trucks, loaders, concrete mixers, helmets	SODA (19,846), MOCS (23,404), Excavator Pose Dataset (8380)	73.2 (detect), 63.7 (segment), 99.5 (pose)	First comprehensive multi-task framework; Unity-generated virtual data	Lower accuracy vs. specialized models; tracking inconsistencies
[30]	Heavy construction equipment	Custom CNN	Sequential CNN architecture	12 classes: Excavator, dump truck, concrete mixer machine/truck, asphalt roller, boom lift, forklift, loader, motor grader, pile driving machine, scissor lift, telescopic handler	Self-collected (10,846 images)	Not reported (P: 0.71–0.87, R: 0.73–0.86, F1: 0.75–0.86)	Most extensive category coverage (12 equipment types)	No multi-site validation; class imbalance
[27]	Multi-object segmentation	YOLOv11x-Seg	C3K2 modules; C2PSA attention	13 classes: Bulldozer, concrete mixer, crane, excavator, hanging head, loader, other vehicle, pile driving, pump truck, roller, static crane, truck, worker	SODA + MOCS combined (40,659 images)	80.8	First YOLOv11-Seg application; 13-category coverage	Uneven class performance; no safety integration
[32]	Hierarchical safety classification	SRGAN + RT-DETR-X + DINOv2	Super-resolution; 3-level hierarchical classification	5 classes: Backhoe, dump truck, excavator, bulldozer, wheel loader	YouTube videos (44,295 images)	95.0	First super-resolution cascade learning; hierarchical safety status	High computational demands (256 GB RAM); limited to 5 classes
[29]	Heavy equipment	YOLOv10 + Swin Transformer	Transformer backbone integration; dual-label assignment	5 classes: Loader, dump truck, excavator, Remicon, crane	Korea construction sites (21,772 images)	89.72	First YOLOv10-transformer integration; real-time 35 FPS	Domain adaptation challenges; Korea-only validation

Table 1. *Cont.*

Study	Detection Target	Base Architecture	Key Technical Innovations	Classes	Dataset (Size)	mAP@0.5 (%)	Notable Contribution	Limitations
[10]	Worker–equipment collision prevention	MSP-YOLO (YOLOv12n-based)	Small Object Enhancement Pyramid; Multi-axis Gated Attention	7 classes: Workers, excavators, loaders, dump trucks, mobile cranes, rollers, bulldozers	Fujian Province, China (22,798 images)	82.5	Multi-scale small object detection; open-source	Computational overhead; dataset not public
[4]	Scaffolding safety	YOLOv12 + RealSense D455	3D depth reasoning; temporal analysis; regulatory compliance	5 classes: Missing guardrail, mobile scaffold—no outrigger, mobile scaffold—outrigger, worker with helmet, worker without helmet, z-Tag	4868 images	97.9	Only scaffolding structural safety study; 3D depth integration	Controlled environment only; no equipment integration
[62]	Comprehensive safety monitoring	Feature Fusion (InceptionV3 + Mobile NetV2)	Model truncation; multi-scale fusion; SE attention	10 classes: Machinery, vehicle, hardhat, no hardhat, safety vest, no safety vest, person, mask, no mask, safety cone	5780 images	81–90	Negative class detection (no hardhat, no vest)	High computational requirements; untested deployment
[60]	Drone-based monitoring	GS-LinYOLOv10	GSConv; Linformer attention; IoT sensor fusion	16 classes: Excavator, wheel loader, dump truck, mixer truck, pump car, scissor lift, tower crane, aerial lift truck, person, hardhat, safety vest, bicycle, car, motorcycle, bus, truck	COCO (200,000+) + Construction Safety (2801)	89.4–90.3	Highest accuracy + fastest inference (92.41 FPS); drone deployment	Small object limitations; no structural safety detection
[61]	Mobile resource tracking	YOLOv8 + PANet	PANet features; twice-association tracking; Enhanced Kalman Filter	13 classes: Workers, dump trucks, excavators, concrete mixers, concrete pump trucks, compactors, wheel loaders, tank trucks, forklifts, semitrucks, sprinklers, crawler cranes, tower cranes	11,039 images	Not reported (Tracking: MOTA 96.49%, IDF1 95.63%, HOTA 90.20%)	First 13-category tracking; fastest real-time (123.6 FPS)	2D tracking only; prolonged occlusion failures
This Study (2025)	Equipment + structural safety	YOLOv11n + 4 variants	Systematic backbone–neck comparison (HFE, FPN, C2F, Identity); dual-dataset design	13 classes: 8 equipment (backhoe, concrete mixer, diesel generator, dump truck, excavator, manpower, tower hoist, tractor) + 5 safety (missing guardrail, mobile scaffold variants, helmet compliance)	ConstructSight (2195 images) + SafeGuard (12,645 images)	92.4–96.8	Distribution-dependent effectiveness quantified; class imbalance impact on architecture selection	Worker detection challenges; Layer 10 focus only

References

1. Shin, Y.; Choi, Y.; Won, J.; Hong, T.; Koo, C. A new benchmark model for the automated detection and classification of a wide range of heavy construction equipment. *J. Manag. Eng.* **2024**, *40*, 04023069. [CrossRef]
2. Li, X.; Ji, H. Enhanced safety helmet detection through optimized YOLO11: Addressing complex scenarios and lightweight design. *J. Real-Time Image Process.* **2025**, *22*, 128. [CrossRef]
3. Huang, K.; Abisado, M.B. Lightweight construction safety behavior detection model based on improved YOLOv8. *Discov. Appl. Sci.* **2025**, *7*, 326. [CrossRef]
4. Abbas, M.S.; Hussain, R.; Zaidi, S.F.A.; Lee, D.; Park, C. Computer Vision-Based Safety Monitoring of Mobile Scaffolding Integrating Depth Sensors. *Buildings* **2025**, *15*, 2147. [CrossRef]
5. Hou, L.; Chen, C.; Wang, S.; Wu, Y.; Chen, X. Multi-Object Detection Method in Construction Machinery Swarm Operations Based on the Improved YOLOv4 Model. *Sensors* **2022**, *22*, 7294. [CrossRef]
6. Shanti, M.Z.; An, B.; Yeun, C.Y.; Cho, C.S.; Damiani, E.; Kim, T.Y. Enhancing Worker Safety at Heights: A Deep Learning Model for Detecting Helmets and Harnesses using DETR Architecture. *IEEE Access* **2025**, *13*, 12345–12356. [CrossRef]
7. Sun, L.; Li, H.; Wang, L. HWD-YOLO: A New Vision-Based Helmet Wearing Detection Method. *Comput. Mater. Contin.* **2024**, *80*, 4543–4560. [CrossRef]
8. Xiao, B.; Kang, S.C. Development of an image data set of construction machines for deep learning object detection. *J. Comput. Civ. Eng.* **2021**, *35*, 05020005. [CrossRef]
9. Kim, S.; Hong, S.H.; Kim, H.; Lee, M.; Hwang, S. Small object detection (SOD) system for comprehensive construction site safety monitoring. *Autom. Constr.* **2023**, *156*, 105103. [CrossRef]
10. Zhang, Y.; Lin, C.; Chen, G. Efficient multi-scale detection of construction workers and vehicles based on deep learning. *J. Real-Time Image Process.* **2025**, *22*, 127. [CrossRef]
11. Rahimian, F.P.; Seyedzadeh, S.; Oliver, S.; Rodriguez, S.; Dawood, N. On-demand monitoring of construction projects through a game-like hybrid application of BIM and machine learning. *Autom. Constr.* **2020**, *110*, 103012. [CrossRef]
12. *Data Subjects, Digital Surveillance, AI and the Future of Work: Study*; European Parliament: Luxembourg, 2020. [CrossRef]
13. Abraha, H.H. A pragmatic compromise? The role of Article 88 GDPR in upholding privacy in the workplace. *Int. Data Priv. Law* **2022**, *12*, 276–296. [CrossRef]
14. Saif, W.; Williams, T.; Wong, C.; Dobos, J.; Martinez, P.; Kassem, M. Digital Twin for Safety on Construction Sites: A Real-time Risk Monitoring System Combining Wearable Sensors and 4D BIM. In Proceedings of the 2024 European Conference on Computing in Construction, Chania, Greece, 4–7 July 2024. [CrossRef]
15. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
16. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.K.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
17. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]
18. Fang, Y.; Ma, Y.; Zhang, X.; Wang, Y. Enhanced YOLOv5 algorithm for helmet wearing detection via combining bi-directional feature pyramid, attention mechanism and transfer learning. *Multimed. Tools Appl.* **2023**, *82*, 28617–28641. [CrossRef]
19. Liu, Y.; Zhang, J.; Shi, L.; Huang, M.; Lin, L.; Zhu, L.; Wang, Z.; Zhang, C. Detection method of the seat belt for workers at height based on UAV image and YOLO algorithm. *Array* **2024**, *22*, 100340. [CrossRef]
20. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *119*, 1–88. Available online: http://data.europa.eu/eli/reg/2016/679/oj (accessed on 20 June 2025).
21. Council Directive 89/391/EEC of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work. *Off. J. Eur. Communities* **1989**, *183*, 1–8. Available online: http://data.europa.eu/eli/dir/1989/391/oj (accessed on 20 June 2025).
22. Martínez-Aires, M.D.; López-Alonso, M.; Aguilar-Aguilera, A.; de la Hoz-Torres, M.L.; Costa, N.; Arezes, P. The General Principles of Prevention in the Framework Directive 89/391/EEC of Occupational Risk Prevention for the Building Construction. In *Occupational and Environmental Safety and Health VI*; Studies in Systems, Decision and Control; Springer: Cham, Switzerland, 2025; Volume 230, pp. 293–302. [CrossRef]
23. Council Directive 92/57/EEC of 24 June 1992 on the implementation of minimum safety and health requirements at temporary or mobile construction sites. *Off. J. Eur. Communities* **1992**, *245*, 6–22. Available online: http://data.europa.eu/eli/dir/1992/57/oj (accessed on 20 June 2025).

24. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, *1689*, 1–144. Available online: http://data.europa.eu/eli/reg/2024/1689/oj (accessed on 20 June 2025).
25. Wang, Q.; Liu, H.; Peng, W.; Tian, C.; Li, C. A vision-based approach for detecting occluded objects in construction sites. *Neural Comput. Appl.* **2024**, *36*, 10825–10837. [CrossRef]
26. Liu, L.; Guo, Z.; Liu, Z.; Zhang, Y.; Cai, R.; Hu, X.; Yang, R.; Wang, G. Multi-Task Intelligent Monitoring of Construction Safety Based on Computer Vision. *Buildings* **2024**, *14*, 2429. [CrossRef]
27. He, L.; Zhou, Y.; Liu, L.; Ma, J. Research and Application of YOLOv11-Based Object Segmentation in Intelligent Recognition at Construction Sites. *Buildings* **2024**, *14*, 3777. [CrossRef]
28. Xuehui, A.; Li, Z.; Zuguang, L.; Chengzhi, W.; Pengfei, L.; Zhiwei, L. Dataset and benchmark for detecting moving objects in construction sites. *Autom. Constr.* **2021**, *122*, 103482. [CrossRef]
29. Eum, I.; Kim, J.; Wang, S.; Kim, J. Heavy Equipment Detection on Construction Sites Using You Only Look Once (YOLO-Version 10) with Transformer Architectures. *Appl. Sci.* **2025**, *15*, 2320. [CrossRef]
30. Yamany, M.S.; Elbaz, M.M.; Abdelaty, A.; Elnabwy, M.T. Leveraging convolutional neural networks for efficient classification of heavy construction equipment. *Asian, J. Civ. Eng.* **2024**, *25*, 6007–6019. [CrossRef]
31. Duan, R.; Deng, H.; Tian, M.; Deng, Y.; Lin, J. SODA: A large-scale open site object detection dataset for deep learning in construction. *Autom. Constr.* **2022**, *142*, 104499. [CrossRef]
32. Kim, B.; An, E.J.; Kim, S.; Sri Preethaa, K.R.; Lee, D.E.; Lukacs, R.R. SRGAN-enhanced unsafe operation detection and classification of heavy construction machinery using cascade learning. *Artif. Intell. Rev.* **2024**, *57*, 206. [CrossRef]
33. 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, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
34. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
35. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science, Volume 9905; Springer: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
37. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
38. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018, Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science, Volume 11211; Springer: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
39. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
40. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098. [CrossRef]
41. Xiao, F.; Mao, Y.; Tian, G.; Chen, G.S. Partial-model-based damage identification of long-span steel truss bridge based on stiffness separation method. *Struct. Control Health Monit.* **2024**, *2024*, 5530300. [CrossRef]
42. Xiao, F.; Mao, Y.; Sun, H.; Chen, G.S.; Tian, G. Stiffness separation method for reducing calculation time of truss structure damage identification. *Struct. Control Health Monit.* **2024**, *2024*, 5171542. [CrossRef]
43. Mao, Y.; Xiao, F.; Tian, G.; Xiang, Y. Sensitivity analysis and sensor placement for damage identification of steel truss bridge. *Structures* **2025**, *73*, 108310. [CrossRef]
44. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
45. Saeheaw, T. SC-YOLO: A Real-Time CSP-Based YOLOv11n Variant Optimized with Sophia for Accurate PPE Detection on Construction Sites. *Buildings* **2025**, *15*, 2854. [CrossRef]
46. Alkhammash, E.H. Multi-Classification Using YOLOv11 and Hybrid YOLO11n-MobileNet Models: A Fire Classes Case Study. *Fire* **2025**, *8*, 17. [CrossRef]
47. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to Attend: Convolutional Triplet Attention Module. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 3139–3148. [CrossRef]
48. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A New Backbone that Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391. [CrossRef]

49. AI IN CIVIL. Construction Project Monitoring Dataset. Roboflow Universe. 2024. Available online: https://universe.roboflow.com/ai-in-civil/02.-construction-project-monitoring (accessed on 20 June 2025).
50. CombinedMobileScaffolding. MobileScaffoldingCheck Dataset. Roboflow Universe. 2024. Available online: https://universe.roboflow.com/combinedmobilescaffolding/mobilescaffoldingcheck (accessed on 20 June 2025).
51. Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance problems in object detection: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3388–3415. [CrossRef]
52. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014, Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Lecture Notes in Computer Science, Volume 8693; Springer: Cham, Switzerland, 2014; pp. 740–755. [CrossRef]
53. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
54. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
55. Lyu, Y.; Yang, X.; Guan, A.; Wang, J.; Dai, L. Construction personnel dress code detection based on YOLO framework. *CAAI Trans. Intell. Technol.* **2024**, *9*, 709–721. [CrossRef]
56. Yipeng, L.; Junwu, W. Personal protective equipment detection for construction workers: A novel dataset and enhanced YOLOv5 approach. *IEEE Access* **2024**, *12*, 47338–47358. [CrossRef]
57. Yang, G.; Hong, X.; Sheng, Y.; Sun, L. YOLO-Helmet: A novel algorithm for detecting dense small safety helmets in construction scenes. *IEEE Access* **2024**, *12*, 107170–107180. [CrossRef]
58. Chen, J.; Zhu, J.; Li, Z.; Yang, X. YOLOv7-WFD: A novel convolutional neural network model for helmet detection in High-Risk workplaces. *IEEE Access* **2023**, *11*, 113580–113592. [CrossRef]
59. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
60. Song, Y.; Chen, Z.; Yang, H.; Liao, J. GS-LinYOLOv10: A drone-based model for real-time construction site safety monitoring. *Alex. Eng. J.* **2025**, *120*, 62–73. [CrossRef]
61. Deng, R.; Wang, K.; Mao, Y. Real-Time Monitoring of Mobile Construction Resources Based on Multiple Object Tracking. *J. Comput. Civ. Eng.* **2025**, *39*, 04025059. [CrossRef]
62. Sharifzada, H.; Wang, Y.; Sadat, S.I.; Javed, H.; Akhunzada, K.; Javed, S.; Khan, S. An Image-Based Intelligent System for Addressing Risk in Construction Site Safety Monitoring Within Civil Engineering Projects. *Buildings* **2025**, *15*, 1362. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.