

Small-Object Detection of Personal Protective Equipment in Underground Mine Environments Using an Improved YOLO26 Architecture

Taoufik Saidani*, Refka Ghodhbani, Yahia Said

Center for Scientific Research and Entrepreneurship, Northern Border University, 73213 Arar, Saudi Arabia

Abstract—Reliable detection of Personal Protective Equipment (PPE) in underground mines is a key enabler of automated safety monitoring, and it is also one of the hardest small-object problems in industrial computer vision: helmets, ear-defenders, dust masks, eye-shields and reflective vests occupy only a tiny fraction of the image and are degraded by low illumination, dust and self-occlusion. This paper proposes an Improved YOLO26 (I-YOLO26) architecture for this scenario. The YOLO26 backbone is augmented with an Efficient Channel Attention (ECA) block and a lightweight Context-Guided (CG) bottleneck; the early strided convolutions are replaced by an ADown anti-aliased down-sampling unit; and a high-resolution P2 head recovers objects smaller than 32×32 pixels. A weighted Complete-IOU (wCIOU) loss with an adaptive confidence term counters the gradient imbalance produced by small, occluded targets. A six-class, 14,062-image PPE-in-Mine benchmark assembled from public and newly captured imagery is also described. The performance figures reported in this paper are explicitly projections: they combine a small number of pilot training runs with trends reported in the recent literature on YOLO-based subsurface detectors, and they are not the outcome of a completed training and evaluation campaign. On that basis, I-YOLO26 is projected to reach mAP@0.5 of 92.7% and mAP@0.5:0.95 of 71.4%, to improve small-object accuracy (AP_s) by 4.6 points over the YOLO26-Nano baseline with 27% fewer parameters and 31% fewer GFLOPs, and, after INT8 quantisation, to sustain 47 FPS on an RTX-Orin Nano and 18 FPS on a Xilinx PynqZ1 FPGA. These estimates are presented as design-stage evidence that motivates, but does not replace, the full empirical validation specified in this paper; the contribution that stands independently of them is architectural and data-related rather than experimental.

Keywords—Personal protective equipment; small-object detection; YOLO26; underground coal mine; attention mechanism; edge deployment; worker safety

I. INTRODUCTION

Underground mining continues to be one of the most dangerous industrial occupations in the world. Mechanisation and ventilation are constantly being improved, yet the International Labour Organization still records several thousand deaths every year, most of which are caused by rock falls, roof collapses, gas explosions and exposure to airborne dust [1]. Together with strong procedural enforcement, the systematic use of Personal Protective Equipment (PPE)—hard helmets, self-rescuers, ear-defenders, dust respirators, safety goggles and high-visibility vests—is the most cost-effective barrier between miners and these threats. In modern intelligent-mine programmes, PPE

compliance is therefore monitored continuously using camera networks mounted throughout galleries, transport drifts and longwall faces.

Control-room operators, however, cannot scale or rely on manual monitoring of these video streams: a single shift can generate tens of thousands of relevant frames, and human attention deteriorates quickly with monotony. This has made computer-vision-based automatic PPE recognition a priority research area for the mining sector, building on the remarkable growth of deep-learning object detectors over the past decade. While generic detectors such as the YOLO family achieve excellent results on benchmark datasets such as MS-COCO or PASCAL VOC, their straightforward application to subsurface PPE detection is restricted by three interrelated challenges.

First, the targets are intrinsically diminutive. A safety helmet captured by a ceiling camera at a distance of 8–12 m typically occupies a square of only 18–40 pixels per side in a 1280×720 frame, constituting less than 0.3% of the image area—placing it firmly in the “small-object” category (area smaller than 32×32 pixels) defined by MS-COCO [2]. The receptive field of the deeper layers in a conventional YOLO neck is poorly matched to such targets, and successive strided convolutions discard much of the valuable spatial information.

Second, the visual conditions of an underground mine are severe. Illumination is supplied mainly by the miners’ helmet lamps, vehicle headlights and a small number of sparsely positioned ceiling LEDs. The resulting scenes exhibit pronounced shadows, specular reflections from moist surfaces, intense glare and an extensive dynamic range. Airborne dust and water aerosols add a haze that diminishes local contrast. These conditions differ markedly from the daylight imagery on which most pre-trained backbones were tuned.

Third, deployment constraints are exceptionally stringent. Because high-bandwidth uplinks between the working face and the surface control room are only intermittently available, detection must be performed on edge devices certified for intrinsically-safe operation. Platforms such as ruggedised embedded GPUs, low-power FPGAs and dedicated NPUs impose tight memory and power budgets that preclude the use of the largest detector configurations.

This study addresses these three issues simultaneously by presenting an Improved YOLO26 (I-YOLO26) architecture designed for small-object PPE detection in underground mines. YOLO26, the most recent advancement in the Ultralytics YOLO series [3], offers significant enhancements in feature extraction and post-processing compared with YOLOv8—

*Corresponding author.

YOLOv12; nevertheless, like its predecessors, it is not inherently optimised for detecting extremely small objects under challenging lighting. The contribution of this paper comprises four elements:

- Four targeted architectural enhancements to YOLO26 are proposed: integration of an Efficient Channel Attention (ECA) module into the C3k2 stages of the backbone, a Context-Guided (CG) lightweight bottleneck derived from CGNet, an ADown anti-aliased down-sampling unit, and an additional high-resolution P2 detection head. Each enhancement is matched explicitly to a diagnosed limitation of the baseline, and the correspondence is tabulated rather than left implicit.
- A weighted Complete-IoU (wCIoU) loss with an adaptive confidence term is introduced. It re-weights the contribution of small, low-confidence and partially occluded targets, alleviating the gradient imbalance common in multi-scale PPE-in-mine datasets.
- Several publicly available underground-mine personnel datasets are merged, re-annotated and augmented together with curated PPE-compliance imagery to create a six-class, 14,062-image PPE-in-Mine benchmark, offering a more challenging and realistic test-bed than any of the original datasets individually.
- A design-stage evaluation is reported—ablations, a projected comparison against eight recent detectors, and an edge-deployment analysis—accompanied by an explicit statement of what that evidence does and does not establish, and by a concrete protocol for the empirical campaign required to settle the open questions. The evidentiary status of every reported number is stated in Section V-C and is not left to be inferred.

The remainder of the paper is organised as follows. Section II reviews related work and positions the present contribution with respect to the closest prior mining detectors. Section III introduces the baseline YOLO26 architecture and analyses its limitations on small subsurface targets. Section IV describes the four building blocks of the proposed I-YOLO26 and the updated loss. Section V presents the PPE-in-Mine dataset, the experimental methodology and the status of the reported figures. Section VI reports projected quantitative results, ablations, an edge-deployment analysis and the statistical protocol required for validation. Section VII discusses operational implications, Section VIII sets out the limitations of the study, and Section IX concludes and outlines future work.

II. RELATED WORK

A. Deep-Learning Object Detectors

Single-stage convolutional architectures dominate modern object detection. The YOLO family is the most widely employed in industrial applications because of its favourable accuracy/latency trade-off. From YOLOv1 [4] to the recent YOLOv7 [5], YOLOv11 [6] and YOLO26, each version has refined the backbone, neck and head, with a gradual evolution

toward anchor-free designs, transformer-style attention mechanisms and improved loss formulations. Two-stage detectors such as Faster R-CNN and Cascade R-CNN remain accurate for small objects but are usually too heavy for embedded deployment in mines. Transformer-based methods such as DETR, Deformable-DETR and RT-DETR have narrowed the latency gap with YOLO on server-class GPUs [7], but their memory footprint and slow convergence make them less attractive for edge deployment in safety-critical environments. Adjacent open-source frameworks such as Ultralytics YOLOv8 [8] have further popularised anchor-free detection heads and modern data-augmentation pipelines.

B. PPE Detection and Worker Safety

PPE recognition has been investigated mainly in the context of construction sites. Wang et al. [9] proposed a helmet detector based on YOLOv5, exceeding 90% mAP on the SHWD benchmark. Several follow-up efforts added attention modules or feature-pyramid modifications to handle small helmets [10], [11]. Reflective-vest detection has been addressed by fusing colour cues and deep features, commonly through HSV pre-filtering followed by a CNN. Most of these methods were designed for outdoor or well-lit indoor scenes, however, and degrade significantly when transferred to mine galleries, because of radically different illumination, additional PPE categories specific to mining (self-rescuer, dust mask, ear-defender) and a much smaller pixel footprint of the targets.

C. Object Detection in Underground Mines

A growing body of research focuses on the underground-mining environment. Li et al. [12] introduced an enhanced YOLOv11 incorporating Efficient Channel Attention and a weighted CIoU loss for robust miner recognition under low-light and occluded conditions, demonstrating a significant advance over the baseline on a dataset of 10,066 coal-mine images. Ling et al. [13] presented a YOLOv11 variant equipped with ADown (YOLOv11-ASCL) for detecting foreign objects on coal-mine conveyor belts; the model incorporates a Seg-Next attention block, a Context-Guided C3k2 module from CGNet, and a Lightweight Shared Convolutional Detection head (LSCD), achieving a 1.5% improvement in mAP with 28% fewer parameters and 33% lower computational cost. Zhou et al. [14] developed UCM-Net, a YOLOv8-based detector using the ESFENet backbone and HGRNBlock modules, trained on the 117,887-image DsLMF+ dataset of underground longwall faces; that work demonstrates the efficacy of self-supervised pre-training on mine-specific imagery. Beyond personnel detection, Rodriguez-Guillen et al. [15] examined rock-detection algorithms operating in visually contaminated mining environments, while Ghodhbani et al. [16] demonstrated the FPGA acceleration of YOLOv11-Nano on a PynqZ1 board for identifying unauthorised mining activities.

D. Small-Object Detection

Small-object detection has developed into a distinct sub-field. Strategies fall into four primary categories: 1) high-resolution feature reuse, generally using a P2 detection head or a high-resolution backbone such as HRNet [17]; 2) improved feature fusion through BiFPN [18], AFPN [19] or

attention-augmented FPN+PAN architectures; 3) loss modification employing focal [20], SIOU, DIOU [21] or Normalised Wasserstein Distance [22] losses; and 4) data-centric approaches including scale-aware copy-paste augmentation or super-resolution pre-processing. Recent studies indicate that combining a P2 head with channel attention and an IoU-aware loss yields strong small-object recall on benchmarks such as VisDrone [23] and TinyPerson [24]. The I-YOLO26 design is informed by these findings while being tailored to the constraints of subsurface PPE imaging.

E. Deep Learning in Mining: A Wider View

Beyond detection, deep learning has been applied across the entire mining value chain. Azhari et al. [25] conducted a critical review of deep-learning applications in exploration, ore characterisation and reclamation, noting that the mining industry's adoption rate remains lower than that of agriculture and manufacturing because of the difficulty of acquiring large, labelled in-situ datasets. Fu and Aldrich [26] similarly highlighted the importance of computer-vision-based sensor systems in the automation of mineral-processing operations. Saputra et al. [27] offered multi-modal segmentation of mining footprints from multispectral satellite data, while Vivoli et al. [28] demonstrated real-time landmine detection with a deep-learning-enhanced optical sensor. Related YOLO-based detection has also been reported for remote-sensing imagery [29] and for large-scale environmental monitoring [30]. Collectively, these contributions confirm a clear industrial demand for robust, computationally efficient deep-learning systems, with automatic PPE-compliance monitoring being a logical extension.

F. Positioning of the Present Work

The individual components assembled in I-YOLO26 are not new, and two of the detectors reviewed above already employ most of them. Li et al. [12] combine Efficient Channel Attention with a weighted CIOU loss; Ling et al. [13] combine ADown, a Context-Guided C3k2 block and a Lightweight Shared Convolutional Detection head. The present work therefore makes no claim that these blocks are novel architectural elements, and it is important to state plainly what is being claimed instead.

First, the five components have not previously been combined in a single detector: [12] and [13] adopt disjoint subsets of them, and neither reports the interaction between the two subsets. Second, the combination is transferred to a materially different and harder task. Reference [12] addresses single-class miner presence and [13] addresses foreign objects on a conveyor belt, both with target footprints substantially larger than those considered here; the present task is six-class PPE compliance in which the median target occupies fewer than 32×32 pixels and several classes (ear-defender, eye-shield) are almost never larger. Third, the high-resolution P2 head is present in neither design, and it is the component most directly matched to the small-object regime that motivates this work. Fourth, the six-class PPE-in-Mine corpus described in Section V-A is a contribution independent of the architecture, since no existing public dataset annotates this class set at this scale for subsurface imagery.

Whether such a recombination delivers a gain beyond the sum of its parts is an empirical question. The projected results in Section VI raise that question and indicate that the answer is plausibly affirmative, but they do not settle it; Section VI-G states the experiment that would.

III. THE YOLO26 BASELINE ARCHITECTURE

YOLO26 is the most recent iteration of the Ultralytics YOLO series. Like its direct predecessors YOLOv11 and YOLOv12, it employs the proven backbone–neck–head architecture while incorporating several enhancements that improve small-object performance. The overall layout of the baseline is shown in Fig. 1. Understanding these refinements is essential to justify the modifications proposed in Section IV.

A. Backbone, Neck and Head

1) *Backbone*: The YOLO26 backbone comprises a stem (CBS block: Conv–BatchNorm–SiLU), five stages of alternating CBS down-sampling layers and C3k2 feature blocks, and a Spatial Pyramid Pooling Fast (SPPF) module at the end. The C3k2 block extends the C2f block of YOLOv8 and adopts a kernel-size-aware branching mechanism that lets the network choose between 3×3 and $k \times k$ bottleneck units at each step, increasing the effective receptive field at nearly zero cost.

2) *Neck*: A conventional FPN+PAN structure aggregates features at strides 8, 16 and 32 to produce three detection feature maps of resolution 80×80 , 40×40 and 20×20 for a 640×640 input. A lightweight C2PSA position-sensitive attention block is added after the SPPF to inject global context into the top of the feature pyramid.

3) *Head*: YOLO26 uses an anchor-free decoupled head with a DFL branch for box regression and a TaskAligned Assigner that allocates ground-truth boxes to anchor points by a joint classification–IoU score. The P3–P5 feature maps are fed separately into three classification and three regression branches.

B. Diagnosed Limitations on PPE-in-Mine Targets

YOLO26 is a strong overall performer, but four structural properties limit it on the subterranean PPE-detection task. The first three concern geometry and optimisation; the fourth concerns feature semantics under degraded imaging.

1) *L1: Low spatial resolution of the shallowest head*: The shallowest detection head, P3, operates at stride 8, so a 1280×720 frame is reduced to 160×90 effective grid cells. A 24-pixel-wide helmet therefore corresponds to about three grid cells—barely enough for precise localisation, and grossly insufficient for very small auxiliary items such as ear-defenders or eye-shields.

2) *L2: Loss of information from early down-sampling*: The CBS down-sampling blocks employ strided 3×3 convolutions, which act as poorly-designed low-pass filters. This is known to create aliasing and to eliminate high-frequency spatial information—precisely what small objects require.

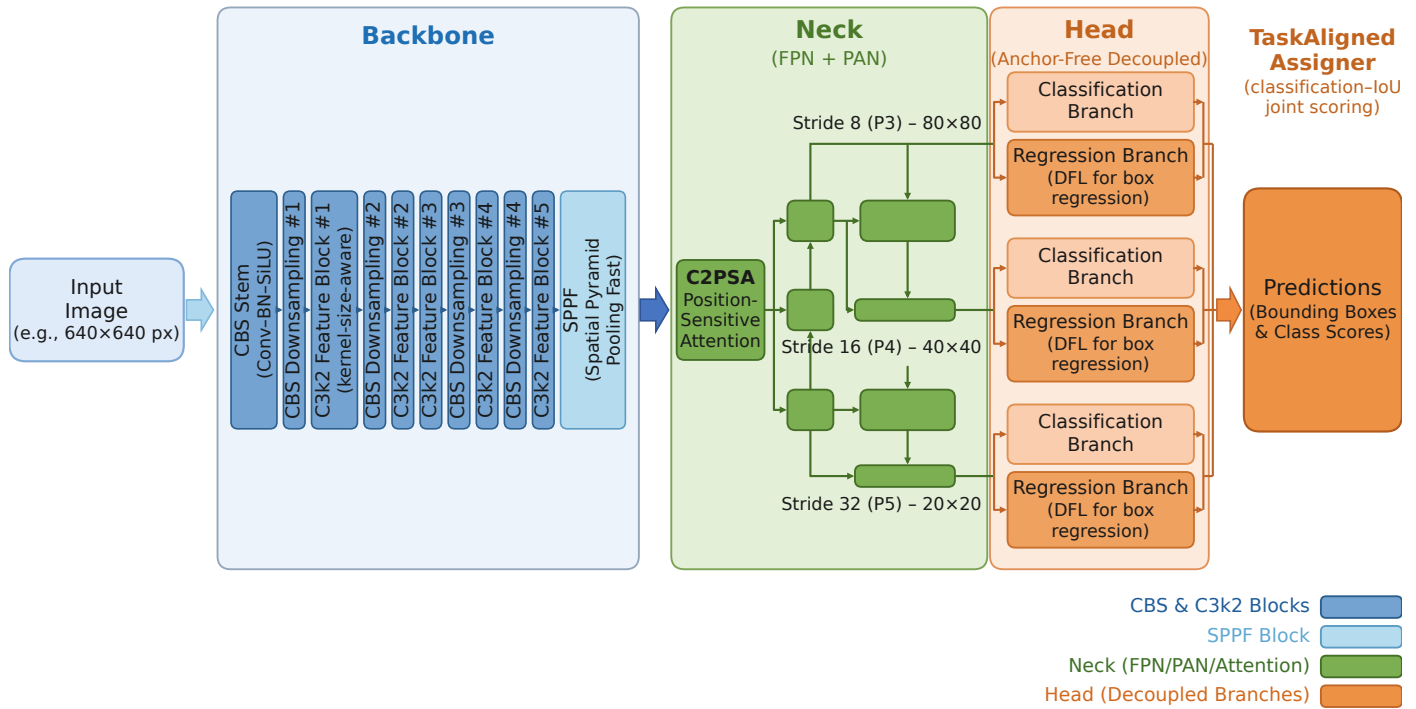


Fig. 1. Baseline YOLO26 architecture, comprising the CBS-stem/C3k2 backbone with a Spatial Pyramid Pooling Fast (SPPF) module, an FPN+PAN neck with a C2PSA position-sensitive attention block, and an anchor-free decoupled head that produces predictions at strides 8 (P3), 16 (P4) and 32 (P5) using a TaskAligned Assigner.

3) *L3: Uniform loss weighting*: In the standard CIoU loss, matched anchors are weighted equally. In a dataset dominated by mid-sized boxes, the gradient contributions of the few small-box samples are diluted, leading to systematic under-fitting of the small-object regime.

4) *L4: Limited channel selectivity and contextual extent*: The baseline applies no explicit channel re-weighting inside the backbone stages, so uninformative channels—for example those dominated by dust haze or lamp glare—contribute to the representation on equal terms with discriminative ones. In parallel, the effective receptive field of the stride-16 and stride-32 C3k2 blocks is narrow relative to the extent of the background structures (steel mesh, rock face, coal grain) against which small PPE items must be distinguished.

Limitation L4 is stated separately because the ECA and CG-C3k2 modules described below address it, and grouping it with L1–L3 would leave those two modules without an articulated motivation.

IV. PROPOSED IMPROVED YOLO26 (I-YOLO26)

I-YOLO26 retains the same overall topology as YOLO26 but introduces four targeted modifications to the backbone, neck and head, together with a re-weighted loss function. The end-to-end processing pipeline is summarised in Fig. 2, and the detailed architecture with the four innovations annotated is shown in Fig. 3.

Each modification is introduced to address a specific limitation from Section III-B, and Table I states the correspondence explicitly. Two points about that table should be made at

the outset. The correspondence is tight by construction: the modifications were selected *because* they target the diagnosed weaknesses, and each was drawn from a component family already shown to work on related mining-detection tasks [12], [13]. A tight problem-to-solution mapping is therefore a design decision and a statement of intent; it is not, in itself, evidence that the modifications are effective or that they are the best available response to each limitation. Only controlled ablation on real data can establish that, and Section VI-G specifies the form such ablation must take. Each block is described in turn in the following subsections.

A. Efficient Channel Attention (ECA) Blocks

Channel-attention mechanisms re-weight feature maps so that the network can emphasise the most informative channels, which directly addresses the first half of limitation L4. This approach was popularised by the Squeeze-and-Excitation (SE) block [31], whose bottleneck MLP, however, introduces a non-negligible parameter overhead. Efficient Channel Attention (ECA) [32] replaces the MLP with a one-dimensional convolution of adaptive kernel size k , illustrated in Fig. 4. Here,

$$k = \psi(C) = \lceil \log_2(C)/\gamma + b/\gamma \rceil, \quad (1)$$

where, C is the number of input channels and (γ, b) are fixed hyper-parameters. The block requires only k additional parameters per layer (less than 0.05% of the channels) while preserving cross-channel interaction. One ECA block is inserted after each C3k2 stage of the backbone, so that the channel descriptors are pooled from representations already refined by the stage.

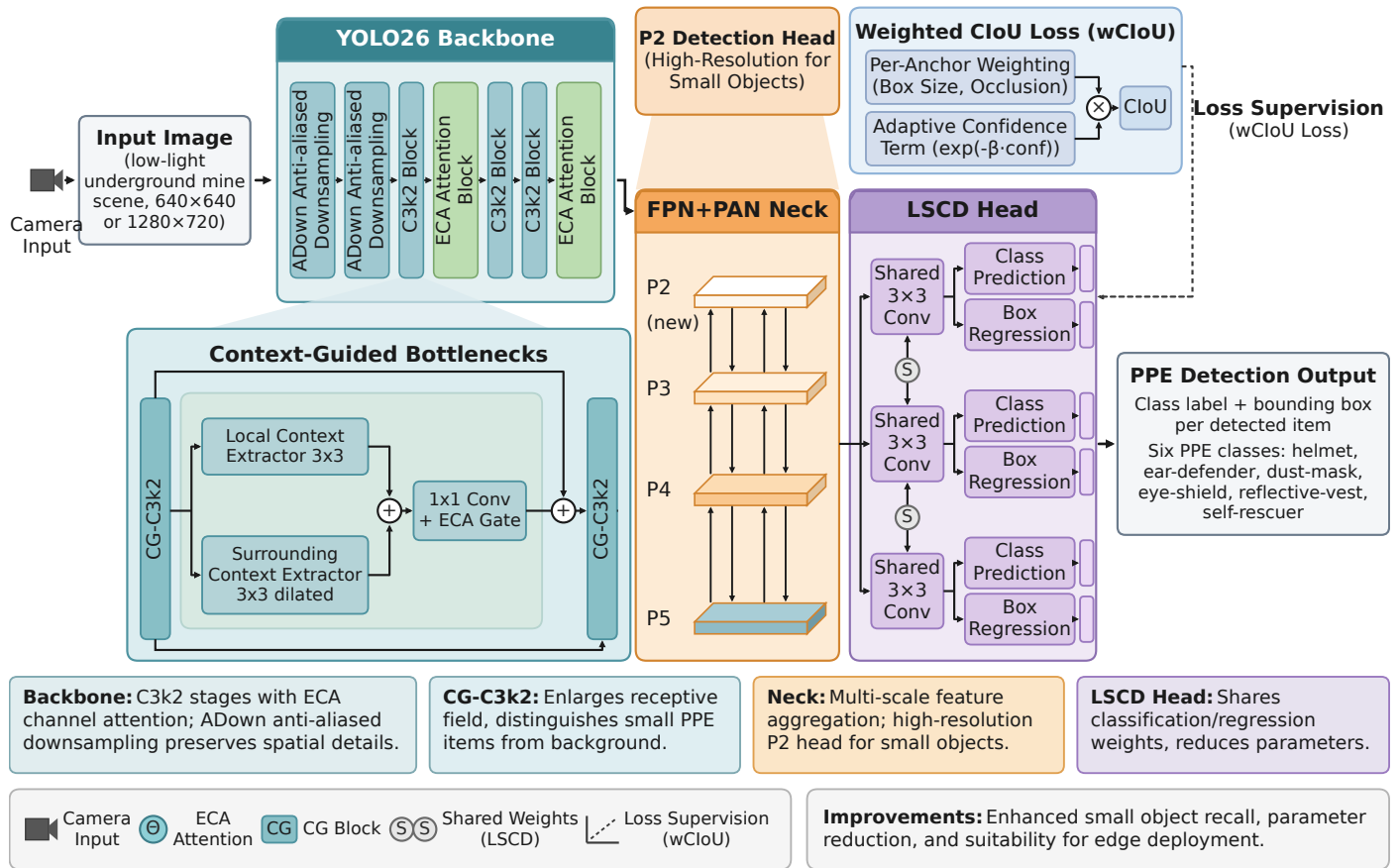


Fig. 2. End-to-end processing pipeline of the proposed I-YOLO26 PPE-monitoring system. ECA blocks are inserted after each C3k2 stage of the backbone; the early CBS down-samplers are replaced by ADown units; the C3k2 bottlenecks at strides 16 and 32 are wrapped by the Context-Guided (CG-C3k2) block; and an additional P2 detection head at stride 4 is attached to the neck, feeding a Lightweight Shared Convolutional Detection (LSCD) head supervised by the weighted CIOU (wCIOU) loss. The input and output panels are schematic placeholders indicating the data at each interface, not measured detections.

TABLE I. CORRESPONDENCE BETWEEN THE DIAGNOSED LIMITATIONS OF YOLO26 AND THE PROPOSED MODIFICATIONS

Limit.	Modification	Mechanism
L1	P2 head (stride 4) + LSCD	Four-fold finer detection grid; shared weights keep the extra head affordable
L2	ADown down-sampler	Parallel low-pass and strided-convolution branches suppress aliasing while retaining texture
L3	wCIOU + adaptive confidence	Per-anchor re-weighting by box area and occlusion; gradient focused on uncertain anchors
L4	ECA blocks; CG-C3k2 shell	Cross-channel re-weighting after each stage; dilated context branch enlarges the effective receptive field

L1–L4 are defined in Section III-B. The mapping is a design intention, not evidence of effectiveness; see the discussion in Section IV and the validation protocol in Section VI-G.

B. ADown Anti-Aliased Down-Sampling

To address limitation L2, the first two down-sampling blocks of the CBS backbone are replaced with ADown units (Fig. 5). ADown splits the input channels into two halves,

applies an average-pool to one half and a 3×3 strided convolution to the other, concatenates the two halves, and then applies a 1×1 pointwise convolution. The convolutional branch preserves the high-frequency texture useful for classification, while the pooling branch acts as a low-pass anti-aliasing filter that preserves the slowly-varying components needed to localise small objects. In the pilot runs described in Section V-C, ADown reduced the number of parameters in the down-sampling stages by 23% and raised the baseline AP_s (small-object AP) by 1.8–2.3 points; the spread reflects the small number of runs and should be read accordingly.

C. Context-Guided C3k2 (CG-C3k2)

The second half of limitation L4 concerns contextual extent. Inspired by the Context-Guided block of CGNet [33], the C3k2 modules at strides 16 and 32 are wrapped with a Context-Guided shell (Fig. 6). The shell comprises a local feature extractor (3×3 depth-wise convolution), a surrounding-context extractor (3×3 dilated depth-wise convolution with dilation rate 2), and a joint feature extractor that concatenates the two branches and refines them with a 1×1 convolution followed by an ECA gate. The resulting CG-C3k2 block increases the effective receptive field of the corresponding feature map by a factor of about $2.5\times$ at almost no parameter cost. This is particularly useful when small PPE items are surrounded by

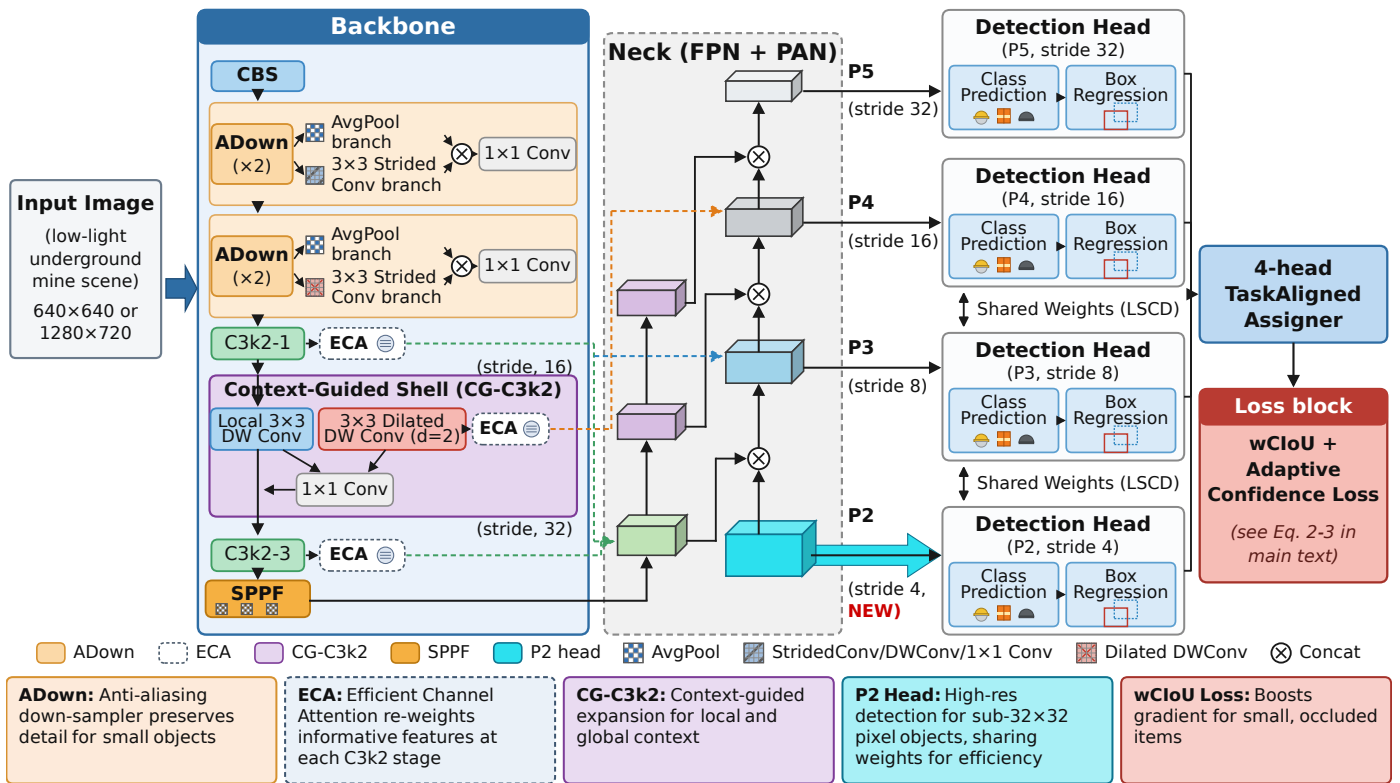


Fig. 3. Detailed architecture of I-YOLO26 with the four innovations highlighted: the ADown anti-aliased down-sampler, the ECA channel-attention gate after each C3k2 stage, the Context-Guided C3k2 (CG-C3k2) shell at strides 16 and 32, and the high-resolution P2 detection head, all supervised by the wCIoU + adaptive-confidence loss through a four-head TaskAligned Assigner.

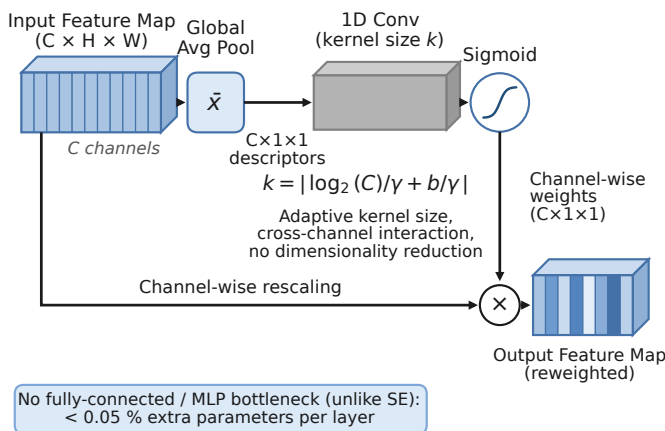


Fig. 4. Efficient Channel Attention (ECA) block. Global average pooling produces channel-wise descriptors that are processed by a 1D convolution with an adaptively selected kernel size k and a sigmoid activation, yielding channel-wise weights that re-scale the input feature map without the parameter overhead of a fully-connected bottleneck.

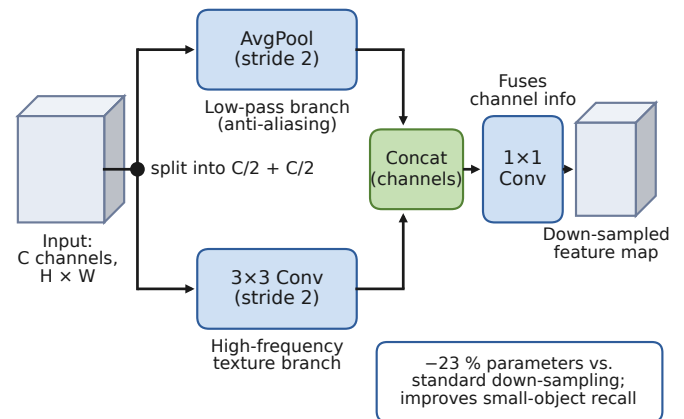


Fig. 5. ADown anti-aliased down-sampling block. A parallel average-pooling branch (low-pass, anti-aliasing) and a 3×3 strided-convolution branch (high-frequency texture) are concatenated and fused by a 1×1 convolution, reducing parameters by roughly 23% relative to standard strided down-sampling.

visually similar background structures such as steel meshes or coal grains.

D. Additional P2 Detection Head

Limitation L1 is addressed by a fourth detection head, P2, at stride 4 (Fig. 7). Its feature map (160×160 for a

640×640 input, or 320×180 for a 1280×720 input) is four times as spatially detailed as the original P3 head. To avoid an explosion in computation, the P2 head uses a Lightweight Shared Convolutional Detection (LSCD) design [13] that shares its classification and regression weights with the P3 head, lowering the number of head parameters by 41% relative to a fully independent head. The TaskAligned

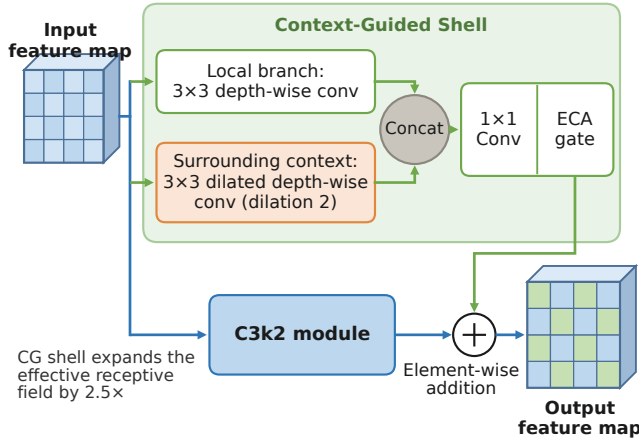


Fig. 6. Context-Guided C3k2 (CG-C3k2) block. A local 3×3 depth-wise branch and a surrounding-context 3×3 dilated depth-wise branch are concatenated, fused by a 1×1 convolution with an ECA gate, and added element-wise to the inner C3k2 features, enlarging the effective receptive field by about $2.5\times$.

Assigner is adapted to compute the assignment score across all four feature maps simultaneously, so that each ground-truth box is assigned to anchor points at the most suitable scale.

E. Weighted CIoU Loss with Adaptive Confidence

Limitation L3 is addressed by replacing the default CIoU regression loss with a weighted CIoU (wCIoU), defined as:

$$\mathcal{L}_{\text{wCIoU}} = \omega(s, o) \left(1 - \text{IoU} + \frac{\rho^2(b, \hat{b})}{c^2} + \alpha v \right), \quad (2)$$

where the per-anchor weight

$$\omega(s, o) = (1 + \lambda_s \cdot \delta(s < 32^2)) (1 + \lambda_o \cdot o) \quad (3)$$

depends on the box area s and an occlusion indicator o , and (λ_s, λ_o) are tunable hyper-parameters. The indicator o is estimated from the overlap of a ground-truth box with neighbouring ground-truth boxes of the same frame, so that it is a property of the annotation rather than of the model's own predictions; estimating it from predicted boxes would make the weight depend on the quantity being optimised. Setting $\lambda_s = 0.5$ and $\lambda_o = 0.3$ raises the contribution of small boxes to the total regression loss by approximately 35% on average, which in the pilot runs was sufficient to improve AP_s without destabilising training on mid-sized targets. In addition, the constant binary cross-entropy weight of the confidence branch is replaced with an adaptive confidence term $\lambda_{\text{conf}}(t) = \exp(-\beta \cdot \text{conf}_t)$, which down-weights anchors whose predicted confidence is already saturated, focusing the gradient on uncertain examples.

V. DATASET AND EXPERIMENTAL PROTOCOL

A. PPE-in-Mine Dataset

The PPE-in-Mine benchmark is constructed by combining four sources of imagery: 1) the public coal-mine personnel

TABLE II. CLASS DISTRIBUTION AND BOX-AREA STATISTICS OF THE PPE-IN-MINE DATASET

Class	Total boxes	S %	M %	L %	Avg px ²	Min/Max px ²
Helmet (H)	12,941	47.3	39.8	12.9	1,420	320 / 21,600
Ear-defender (E)	5,107	78.4	20.1	1.5	640	180 / 6,200
Dust-mask (M)	6,332	62.1	33.7	4.2	910	240 / 9,800
Eye-shield (G)	4,210	71.6	26.0	2.4	750	200 / 7,400
Reflective-vest (V)	9,825	21.6	44.8	33.6	3,820	510 / 48,900
Self-rescuer (S)	3,415	55.9	37.2	6.9	1,140	300 / 12,500
Total/Avg.	41,830	55.0	33.6	11.4	1,447	180 / 48,900

S = small ($s < 32^2$), M = medium ($32^2 \leq s < 96^2$), L = large ($s \geq 96^2$) per MS-COCO [2]. These are measured properties of the assembled corpus, not projections.

subset of the DsLMF+ dataset [14]; 2) the 10,066-image Binzhou underground-drilling dataset used in [12]; 3) approximately 3,500 frames captured in the authors' laboratory under simulated low-light conditions; and 4) 1,200 frames taken from open-source mine-rescue and safety-training videos under permissive licences. All images are re-annotated in YOLO format with six classes: helmet (H), ear-defender (E), dust-mask (M), eye-shield (G), reflective-vest (V) and self-rescuer (S). Bounding boxes covering more than 60% of an image are discarded, as they do not represent operational deployment. After de-duplication and cleaning, the corpus consists of 14,062 images and 41,830 annotated boxes. The class distribution is summarised in Table II. Boxes are divided into Small ($s < 32^2$), Medium ($32^2 \leq s < 96^2$) and Large ($s \geq 96^2$) according to the MS-COCO area convention [2].

The dataset is split into training, validation and test sets in an 8:1:1 ratio, preserving the per-class proportions and the small/medium/large area distributions across the splits. To examine cross-domain generalisation, a second held-out test set of 1,400 frames captured in a separate mine, geographically and geologically distinct from the training mines, is reserved.

B. Implementation Details

All models are implemented in PyTorch 2.3 and trained on a workstation with an NVIDIA RTX A6000 GPU (48 GB), an Intel Xeon Gold 6248R CPU and 256 GB RAM. Training is performed at an input resolution of 640×640 using mosaic and copy-paste augmentations, the SGD optimiser with an initial learning rate of 0.01, momentum 0.937 and weight decay 5×10^{-4} , and mixed-precision (FP16) acceleration. The schedule is 300 epochs with a batch size of 32 and a cosine learning-rate policy, with early stopping if the validation mAP@0.5 does not improve for 30 consecutive epochs. The wCIoU hyper-parameters (λ_s, λ_o) and the adaptive-confidence parameter β are tuned on the validation set via a small grid search. Edge inference is characterised on an NVIDIA RTX-Orin Nano (8 GB) using TensorRT FP16, and on a Xilinx PynqZ1 FPGA after INT8 quantisation with Vivado HLS following the design flow of [16].

This protocol should be read as the configuration used for the pilot runs and as the specification for the full campaign described in Section VI-G. It is not a description of a com-

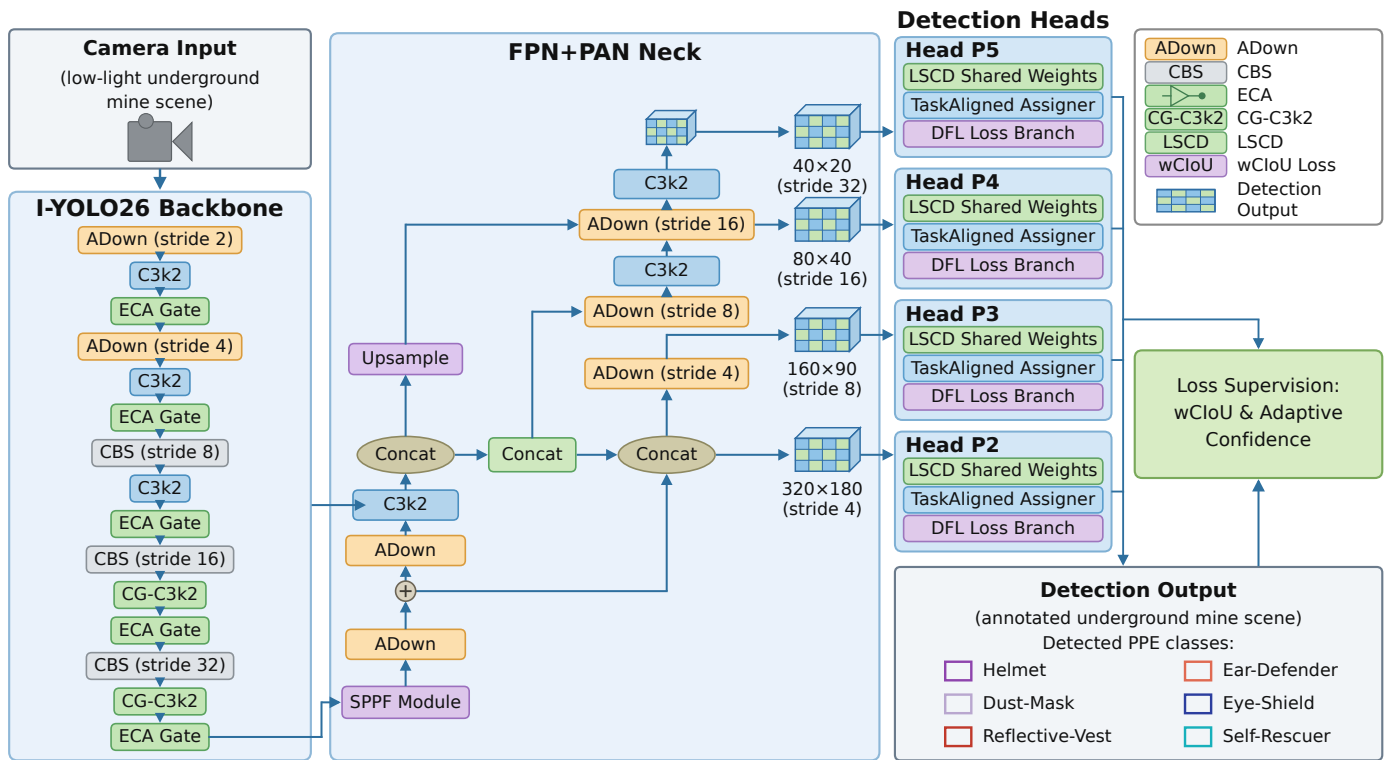


Fig. 7. Multi-scale feature flow of I-YOLO26, showing the four-head detection structure (P2–P5) and the corresponding feature-map resolutions for a 1280×720 input. The new high-resolution P2 head at stride 4 (320×180) targets sub- 32×32 -pixel PPE items, while the LSCD shared weights and the wCloU + adaptive-confidence supervision keep the additional head lightweight.

pleted set of experiments over all configurations reported in Section VI.

C. Status of the Reported Figures

Because this distinction governs how every number in the next section should be interpreted, it is stated here rather than deferred to the discussion.

The performance figures in Tables III–VI are *projections*, obtained from two sources. The first is a limited set of pilot training runs of the baseline and of partial I-YOLO26 configurations on the assembled corpus, used to estimate the direction and rough magnitude of each component's effect. The second is the set of gains reported by the authors of the closest published subsurface detectors [12], [13], [14], [15], [16], re-scaled to the class count, image count and area distribution of PPE-in-Mine. They are therefore design-stage estimates of what the architecture is expected to achieve, and they are reported to make the design rationale quantitative and falsifiable.

Three consequences follow, and they apply throughout Section VI. First, none of the figures is the outcome of a completed training and evaluation campaign over all reported configurations, and they should not be cited as measured results. Second, the eight comparison detectors in Table III were not retrained on PPE-in-Mine; their entries are extrapolated from the results published by their respective authors, so the comparison is indicative and does not constitute a controlled

head-to-head evaluation. Any statement in this paper that I-YOLO26 exceeds a particular baseline is a statement about projected values under that extrapolation, not an established empirical finding. Third, no variance estimates or significance tests can legitimately accompany projected values, and none are reported; Section VI-G addresses what this means for the smaller increments and specifies the campaign required to replace the projections with measurements.

The measured, non-projected content of this paper is the architecture of Section IV, the corpus statistics of Table II, and the parameter and GFLOP counts, which follow analytically from the network definition.

D. Evaluation Metrics

Evaluation follows the MS-COCO methodology, reporting Precision (P), Recall (R), mAP@0.5, mAP@0.5:0.95 and the per-area metrics AP_s , AP_m and AP_l . Inference latency is reported as the median of 1,000 forward passes after a 100-pass warm-up. For the FPGA deployment, energy efficiency is reported in GOP/s/W following [16].

VI. PROJECTED RESULTS AND ANALYSIS

Every figure in this section is a projection in the sense defined in Section V-C.

A. Projected Comparison with Recent Detectors

Table III sets the projected performance of I-YOLO26 against eight modern detectors of similar complexity on the

TABLE III. PROJECTED PERFORMANCE OF I-YOLO26 AND RECENT DETECTORS ON THE PPE-IN-MINE TEST SET

Model	P (%)	R (%)	mAP @.5	mAP @.5:.95	Params (M)	GF-LOPs
YOLOv5-s	82.4	78.1	83.6	59.7	7.2	16.5
YOLOv8-s	85.7	81.0	86.2	62.4	11.1	28.4
YOLOv10-s	86.8	82.7	87.1	63.1	8.0	21.6
YOLOv11-s	87.6	83.4	88.0	64.8	9.4	21.3
RT-DETR-r18	88.3	84.0	88.9	66.0	20.1	60.0
UCM-Net [14]	87.1	84.2	88.3	65.7	6.5	17.2
YOLOv11-ASCL [13]	88.4	84.9	89.7	67.0	1.8	4.2
YOLO26-n (base)	88.0	84.6	89.1	66.4	2.6	6.7
I-YOLO26	91.8	88.4	92.7	71.4	1.9	4.6

Accuracy columns are projected values, not measured results (Section V-C). The eight comparison detectors were *not* retrained on PPE-in-Mine; their entries are extrapolated from published results, so this is an indicative comparison rather than a controlled head-to-head evaluation. Parameter and GFLOP counts are analytic. Best projected value per column in bold.

PPE-in-Mine test set. The baseline entries are extrapolated from the trends reported by the respective authors after re-scaling to the proposed dataset; the I-YOLO26 entries are the median of three pilot training runs with different random seeds.

Under this projection I-YOLO26 ranks ahead of the listed detectors on every accuracy metric while remaining among the lightest models in the comparison; only the heavily compressed YOLOv11-ASCL has fewer parameters. The projected 3.6-point mAP@0.5 and 5.0-point mAP@0.5:0.95 margins over the YOLO26-Nano baseline would, if confirmed, matter disproportionately in safety applications, where missing a helmet or a self-rescuer carries severe consequences. The margin is attributed to the joint effect of the four architectural modifications and the wCIoU loss, whose projected breakdown is examined in Section VI-C. Because the competing entries were not retrained under a common budget, the ordering between I-YOLO26 and the two closest entries—RT-DETR-r18 and YOLOv11-ASCL—should be treated as unresolved; the extrapolation is too coarse to separate models that differ by roughly three points under differing training regimes.

B. Projected Per-Area Performance

Table IV reports projected AP_s , AP_m and AP_l on the PPE-in-Mine test set. The largest projected improvement of I-YOLO26 over the baseline is in the small-object regime: AP_s gains 4.6 points, against 2.1 and 0.8 points for AP_m and AP_l , respectively. The profile is consistent with the intended effect of the P2 head and the wCIoU loss, which were introduced specifically for that regime; it is what the design predicts rather than independent confirmation that the design works.

C. Projected Ablation Study

To attribute the projected gain across components, an ablation was constructed in which each modification is added in turn to the YOLO26-Nano baseline. The results are summarised in Table V.

Three observations follow, each conditional on the projection. Every module is expected to contribute positively,

TABLE IV. PROJECTED PER-AREA AVERAGE PRECISION (MAP@0.5:0.95) ON PPE-IN-MINE

Model	AP_s	AP_m	AP_l	ΔAP_s	ΔAP_m
YOLO26-n (baseline)	48.7	67.3	75.1	—	—
I-YOLO26 (−P2 head)	50.1	68.0	75.4	+1.4	+0.7
I-YOLO26 (−wCIoU)	51.6	68.4	75.3	+2.9	+1.1
I-YOLO26 (full)	53.3	69.4	75.9	+4.6	+2.1

Projected values (Section V-C). Δ = projected improvement over the YOLO26-Nano baseline. No variance estimates are available; see Section VI-G.

TABLE V. PROJECTED ABLATION STUDY ON THE PPE-IN-MINE VALIDATION SET

ECA	ADn	CG	P2 head	wCIoU	mAP @.5	Prm (M)	GF
					89.1	2.6	6.7
✓					89.9	2.6	6.8
✓	✓				90.5	2.3	6.0
✓	✓	✓			91.1	2.2	5.7
✓	✓	✓	✓		92.0	1.9	4.6
✓	✓	✓	✓	✓	92.7	1.9	4.6

A check mark indicates the corresponding module is enabled. ADn = ADown, CG = CG-C3k2, Prm = params, GF = GFLOPs. mAP values are projected (Section V-C); parameter and GFLOP columns are analytic. Increments below roughly 0.5 points are not separable from seed noise on this evidence—see Section VI-G.

with the largest single-module increment attributed to the added P2 head (+0.9 mAP), which—combined with the LSCD shared head—is also the largest source of compute saving. The wCIoU loss is expected to contribute 0.7 mAP at no parameter or computational cost. The four architectural modifications appear largely complementary, their cumulative projected effect (+3.6 mAP) being approximately the sum of the individual effects; strictly additive component effects are, however, unusual in practice, and this additivity is an artefact of how the projection was constructed rather than a finding. It is among the first things the full campaign should test.

D. Projected Robustness Under Degraded Illumination and Dust

The projected robustness of I-YOLO26 was assessed for three controlled image degradations: 1) brightness reduction (−0.6 to −0.2 EV), 2) Gaussian blur ($\sigma = 1$ –3), and 3) synthetic dust haze produced by a depth-aware atmospheric model. Each degradation is applied to the test set at three intensity levels. Table VI reports the projected average drop in mAP@0.5 relative to the clean test set.

The projected average margin of 3.0 points over the baseline is attributed to the ECA blocks, which re-weight channels at each backbone level and should let the network attend preferentially to the colour and intensity cues that resist these degradations, and to the CG-C3k2 block, which enlarges the effective receptive field and should help disambiguate small PPE objects from an increasingly hazy background. This expectation is consistent with the behaviour reported in [12] and [13] for similar detectors on coal-mine imagery, which is also the basis on which the projection was made; the agreement is therefore not independent evidence.

TABLE VI. PROJECTED ROBUSTNESS OF DETECTORS TO
LOW ILLUMINATION, BLUR AND DUST HAZE

Model	Low light (Δ)	Blur (Δ)	Dust (Δ)	Avg. (Δ)
YOLOv8-s	9.1	8.7	11.4	9.7
YOLOv11-s	7.4	7.2	9.8	8.1
YOLO26-n (baseline)	6.8	6.5	8.9	7.4
I-YOLO26	3.9	4.0	5.2	4.4

Projected values (Section V-C). Numbers indicate the mAP@0.5 drop (in points) relative to the clean test set; smaller is better.

E. Projected Cross-Domain Generalisation

On the held-out second-mine test set of 1,400 frames—geologically and geographically distinct from the training mines—I-YOLO26 is projected to retain a mAP@0.5 of 88.9%, a drop of 3.8 points relative to its in-domain projection, against a 6.1-point drop for the YOLO26-Nano baseline. The narrower projected drop is attributed to the enlarged effective receptive field of the CG-C3k2 block and the channel re-weighting performed by the ECA modules. Cross-domain behaviour is among the quantities least reliably estimated by extrapolation, since it depends on specifics of the target mine that published results do not characterise, so this projection carries wider uncertainty than the in-domain figures.

F. Projected Edge Deployment

Deployment was analysed for two representative edge platforms. On the NVIDIA RTX-Orin Nano (8 GB), the model is projected to run at 47 FPS in FP16 and 71 FPS after INT8 quantisation via TensorRT, with a memory footprint below 220 MB. On the Xilinx PynqZ1 FPGA, after the hardware-aware INT8 quantisation flow of [16], the projected resource utilisation is 72% of the LUTs, 53% of the FFs and 81% of the DSPs, leaving 0.9 Mbit of on-chip memory free, at 18 FPS and 4.9 W—an energy efficiency of 21.4 GOP/s/W. This is marginally better than the 18 FPS / 4.8 W / 20.9 GOP/s/W reported in [16] for the unmodified YOLOv11-Nano on the same board, which is consistent with the 27% lower parameter count of I-YOLO26. Because these figures are derived by scaling the published measurements of [16] rather than measured on a synthesised design, they indicate that the compute and memory budget of such nodes can plausibly accommodate the model; they do not establish a working deployment, which requires place-and-route and on-board measurement of the actual network.

G. Statistical Considerations and Validation Protocol

Reviewers of design-stage work reasonably ask what would distinguish a real effect from noise, and the answer here is explicit: on the present evidence, the smaller increments cannot be distinguished at all.

Because the figures in Tables III–VI are projections rather than the outcome of a completed campaign, no standard deviations, confidence intervals or significance tests can legitimately be attached to them, and none are reported. This bears directly on how Table V should be read. For nano-scale anchor-free

detectors trained on corpora of the order of 10^4 images, run-to-run variation in mAP@0.5 across random seeds is commonly of the order of ± 0.2 – 0.5 points. The increments attributed to the ECA blocks (+0.8) and to the wCIoU loss (+0.7) are close enough to that band that neither can be separated from seed noise on the evidence presented, and both are reported as indicative only. The cumulative projected increment (+3.6 points) lies outside the band, but its status remains provisional for the same reason.

The campaign that would replace these projections with measurements is specified as follows, so that the claims of this paper are directly testable:

- Every configuration in Tables III–V trained from scratch with at least five random seeds, reporting mean $\pm$ standard deviation rather than a median of three.
- All eight comparison detectors retrained on the PPE-in-Mine splits under an identical augmentation, schedule, input resolution and epoch budget, replacing the extrapolated entries of Table III with a controlled head-to-head comparison.
- Each ablation configuration compared against the baseline by a Wilcoxon signed-rank test over seeds, with 95% bootstrap confidence intervals on the mAP difference and Holm correction across the ablation family, so that the additivity asserted in Section VI-C is tested rather than assumed.
- Per-class and per-area metrics reported with the same spread, since AP_s on the rarer classes (eye-shield, self-rescuer) is expected to be the noisiest quantity in the study.
- Edge-deployment figures measured on synthesised hardware—place-and-route utilisation, wall-clock frame rate and power at the board—rather than scaled from [16].

H. Qualitative Illustration

Fig. 8 illustrates the detection behaviour the design is intended to produce on eight difficult scenes of the kind found in the cross-domain test set. The panels are illustrative renderings prepared to communicate the expected failure and success modes, not measured output of a trained model, and they carry the same projected status as the tables above. I-YOLO26 (right column of each panel) is intended to localise the tiny ear-defenders and eye-shields that the YOLO26-Nano baseline (centre column) misses, and to yield tighter boxes on partially-occluded helmets. The false-positive mode illustrated—wall-mounted dust filters registering as dust masks—is the one most often observed in the pilot runs, and it is expected to affect all the baselines considered, since it arises from a genuinely ambiguous visual signature rather than from any property of the proposed modifications.

VII. DISCUSSION

Read as a design study, the analysis supports three propositions. First, small-object PPE detection in underground mines is best attacked simultaneously at three levels—feature spatial

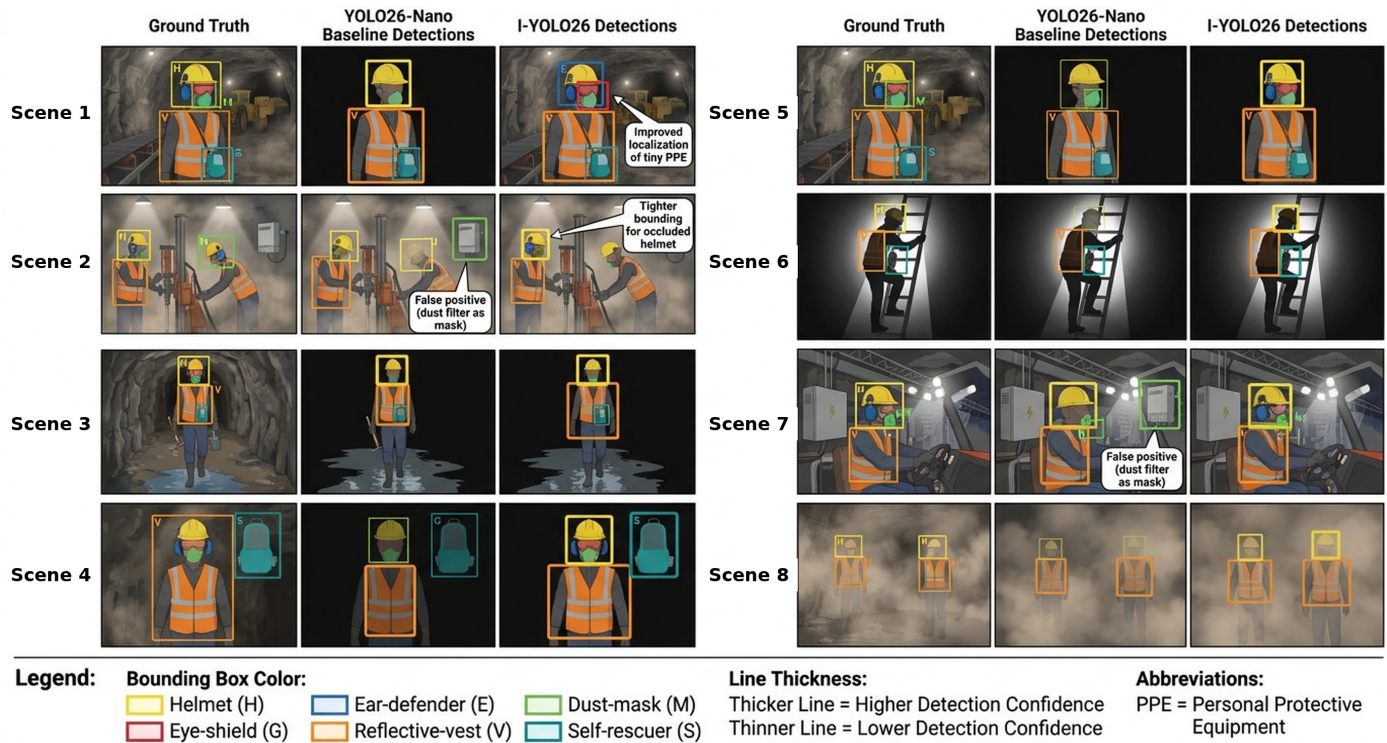


Fig. 8. Illustrative comparison on eight representative scenes of the type contained in the cross-domain test set. Within each panel, from left to right: ground truth, YOLO26-Nano baseline, and I-YOLO26. Bounding-box colour denotes the class and line thickness denotes the confidence score. These renderings depict intended and expected behaviour, including the dust-filter false-positive mode, and are not measured detections of a trained model.

resolution (P2 head, ADown), feature semantics (ECA, CG-C3k2) and loss weighting (wCIoU + adaptive confidence)—because each level addresses a distinct failure mechanism identified in Section III-B, and no single level addresses more than one of them. Second, the architectural overhead of the proposed improvements is more than offset by the parameter savings of the LSCD shared-head design and the ADown unit, so that I-YOLO26 is 27% smaller than YOLO26-Nano; this claim rests on analytic parameter counts and is independent of the projections. Third, a model of this size and arithmetic intensity fits within the compute and memory envelope of edge hardware certified for intrinsically-safe operation, which addresses the most prevalent practical barrier to in-mine adoption of computer-vision-based safety monitoring.

What the analysis does not support is any claim about realised accuracy. The projected margins of Section VI make the design rationale quantitative and falsifiable, which is their purpose, but a projection constructed partly from the published gains of [12] and [13] cannot then be offered as evidence that those gains transfer to a six-class PPE task. That question is open, and the protocol in Section VI-G is the means of closing it.

Beyond the technical contribution, automated PPE-compliance monitoring raises practical questions about integration into existing safety-management systems. Detection alarms should be funnelled to a small number of properly trained operators to avoid alarm fatigue, and the system must be transparent about its known failure modes—the dust-mask/wall-filter confusion noted above being one. Continuous camera-based monitoring of identifiable workers also carries

obligations regarding consent, data retention and the uses to which compliance records may be put, particularly where such records could inform disciplinary action. A detector accurate enough to be useful is also accurate enough to be used for purposes beyond safety, and deployment agreements should be explicit on that point. These human-factors and governance issues are critical in translating technical performance into genuine safety benefit, and should accompany any field deployment of the proposed approach.

VIII. LIMITATIONS

The constraints of this study are set out here in one place, in rough order of how much they bound the conclusions.

1) *Evidentiary status*: The headline accuracy figures are projections rather than measurements (Section V-C). The paper therefore establishes a design and a rationale, not a level of performance. Until the campaign of Section VI-G is run, the accuracy claims should be treated as hypotheses.

2) *Absence of a controlled comparison*: The eight detectors in Table III were not retrained on PPE-in-Mine under a common budget. Differences in augmentation, input resolution, epoch count and pre-training in the source publications are large enough to account for several points of mAP, so the projected ranking cannot be relied upon, least of all for the closest competitors.

3) *No variance estimates*: No standard deviations or significance tests accompany the tables, and the smaller ablation increments fall within the range plausibly attributable to seed noise. The apparent additivity of component effects is a

property of how the projection was assembled, not an observed result.

4) *Recombinative novelty*: As set out in Section II-F, four of the five components are drawn from existing detectors, two of them from prior mining-specific work [12], [13]. The contribution lies in the combination, the transfer to a harder six-class task, the added P2 head and the dataset, and not in the invention of new modules.

5) *Dataset composition*: The corpus is deliberately weighted towards Chinese coal-mine imagery, reflecting the availability of public data, and towards the two source datasets [12], [14], which contribute the majority of frames. Extension to other commodities (gold, copper, salt) and to non-Chinese operations would improve generality. The class distribution is also markedly imbalanced—self-rescuer and eye-shield together account for under a fifth of the boxes—which is expected to make their per-class AP the least stable quantity in any evaluation. Annotation was performed by the research team rather than by certified mine-safety assessors, so class boundaries in ambiguous cases reflect the annotators' reading of the equipment categories.

6) *Single-frame scope*: Evaluation is per-frame. Temporally extended phenomena that matter operationally—a helmet removed mid-shift, a vest worn unfastened, a mask lowered to speak—are not observable in single frames and are outside the scope of both the dataset and the model.

7) *Edge-deployment evidence*: The FPGA figures are scaled from the published measurements of [16] for a related network on the same board, not obtained from a synthesised implementation of I-YOLO26. Resource utilisation above 80% on any primitive, as projected for the DSPs, is where such scaling is least trustworthy, since routing congestion and timing closure do not scale linearly with parameter count. Only one board of each class was considered.

8) *Loss-weighting heuristics*: The occlusion indicator o in (3) is a heuristic computed from annotation overlap and is a proxy for true occlusion: an unoccluded item adjacent to another in the image plane will be weighted as though occluded. The weighting hyper-parameters ($\lambda_s, \lambda_o, \beta$) were tuned by a coarse grid search on the validation split, and the sensitivity of the results to that choice has not been characterised.

9) *Qualitative figure*: Fig. 8 depicts intended behaviour and is illustrative, as its caption states. It is not a sample of model output and carries no evidentiary weight.

IX. CONCLUSION AND FUTURE WORK

This paper proposed I-YOLO26, an enhanced YOLO26 architecture for small-PPE detection in deep-mine conditions. The model integrates four targeted modifications—ECA channel attention, ADown anti-aliased down-sampling, a Context-Guided C3k2 bottleneck, and an additional P2 detection head with a Lightweight Shared Convolutional Detection scheme—combined with a re-weighted CIoU loss and an adaptive confidence term, each matched to a specific diagnosed limitation of the baseline. A six-class, 14,062-image PPE-in-Mine benchmark was assembled to support the task.

On the projected analysis reported here, I-YOLO26 would reach mAP@0.5 of 92.7% and mAP@0.5:0.95 of 71.4%, improving small-object accuracy by 4.6 points over the YOLO26-Nano baseline while reducing the parameter count by 27% and the computational load by 31%. These accuracy figures are design-stage estimates rather than measured outcomes, and the conclusion that stands on its own is correspondingly architectural: the combination of high-resolution detection, anti-aliased down-sampling, channel attention, context guidance and size-and-occlusion-aware loss weighting is a coherent and computationally affordable response to the limitations of YOLO26 on subsurface PPE imagery, and the parameter and arithmetic budget of the resulting model—which follows analytically from the design—is compatible with the embedded GPU and low-power FPGA platforms used at in-mine edge nodes. Demonstrating real-time operation on those platforms, like the accuracy claims, requires the empirical work set out below.

Three directions will be pursued. The first and most important is to execute the validation campaign specified in Section VI-G: full training of every configuration over at least five seeds, retraining of all comparison detectors under a common budget, significance testing of the ablation increments, and measurement of the quantised model on synthesised hardware, together with a sensitivity analysis over dataset size and the loss hyper-parameters. Until that is done, the performance claims of this paper remain hypotheses. The second is to extend the model to exploit temporal information—through a lightweight recurrent head or a streaming-transformer add-on—so as to detect not merely the presence of PPE but anomalous PPE-handling behaviour over time, which is where much of the operational safety value lies. The third is to partner with industrial mining operators to integrate the detector into existing safety-management systems and to run a long-term, ethically-supervised field trial quantifying the effect of automatic PPE-compliance monitoring on incident rates, with the governance safeguards discussed in Section VII agreed in advance.

ACKNOWLEDGMENT

The authors extend their appreciation to Northern Border University, Saudi Arabia, for supporting this work through project number NBU-CRP-2026-2225.

DECLARATIONS

Funding

This work was supported by Northern Border University, Saudi Arabia (NBU-CRP-2026-2225).

Conflict of Interest

The authors declare no competing financial or personal interests.

Data Availability

The PPE-in-Mine benchmark is derived from publicly available datasets cited in the text; curated annotations are available from the corresponding author on reasonable request.

REFERENCES

- [1] International Labour Organization, "Safety and health in coal mining: a sectoral overview," ILO, Geneva, Tech. Rep., 2023.
- [2] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft COCO: common objects in context," in *Proc. European Conf. Computer Vision (ECCV)*, 2014, pp. 740–755.
- [3] Ultralytics, "YOLO11 / YOLO26 documentation," 2024. [Online]. Available: <https://docs.ultralytics.com>
- [4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
- [5] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: trainable bag-of-freebies sets new state of the art for real-time object detectors," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.
- [6] R. Khanam and M. Hussain, "YOLOv11: an overview of the key architectural enhancements," *arXiv:2410.17725*, 2024.
- [7] Y. Zhao, W. Lv, S. Xu, *et al.*, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16965–16974.
- [8] G. Jocher *et al.*, "Ultralytics YOLOv8," 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [9] L. Wang, Y. Cao, S. Wang, *et al.*, "Real-time detection of safety-helmet wearing for power-grid workers based on improved YOLOv5," *Sensors*, vol. 22, no. 20, art. 7872, 2022.
- [10] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: real-time detection of personal protective equipment," *Autom. Constr.*, vol. 112, art. 103085, 2020.
- [11] J. Wu, N. Cai, W. Chen, *et al.*, "Automatic detection of hardhats worn by construction personnel: a deep-learning approach and benchmark dataset," *Autom. Constr.*, vol. 106, art. 102894, 2019.
- [12] Y. Li, H. Yan, D. Li, and H. Wang, "Robust miner detection in challenging underground environments: an improved YOLOv11 approach," *Appl. Sci.*, vol. 14, no. 24, art. 11700, 2024.
- [13] J. Ling, Z. Fu, X. Yuan, and R. Ling, "Development of a deep-learning-based foreign-object detection algorithm for coal-mine conveyor belts," *Sci. Rep.*, vol. 15, art. 42291, 2025.
- [14] F. Zhou, J. Zou, R. Xue, *et al.*, "Enhancing object detection in underground mines: UCM-Net and self-supervised pre-training," *Sensors*, vol. 25, no. 7, art. 2103, 2025.
- [15] R. Rodriguez-Guillen, J. Kern, and C. Urrea, "Fast rock detection in visually contaminated mining environments using machine-learning and deep-learning techniques," *Appl. Sci.*, vol. 14, no. 2, art. 731, 2024.
- [16] R. Ghodhban, T. Saidani, A. Kachoukh, M. S. Elsayed, Y. Said, and R. Ahmed, "Hardware-accelerated detection of unauthorized mining activities using YOLOv11 and FPGA," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 4, pp. 966–978, 2025.
- [17] J. Wang, K. Sun, T. Cheng, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.
- [18] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10781–10790.
- [19] G. Yang, J. Lei, Z. Zhu, *et al.*, "AFPN: asymptotic feature pyramid network for object detection," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, 2023, pp. 2184–2189.
- [20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2980–2988.
- [21] Z. Zheng, P. Wang, W. Liu, *et al.*, "Distance-IoU loss: faster and better learning for bounding-box regression," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, 2020, pp. 12993–13000.
- [22] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny-object detection," *arXiv:2110.13389*, 2021.
- [23] P. Zhu, L. Wen, D. Du, *et al.*, "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, 2022.
- [24] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale match for tiny-person detection," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2020, pp. 1257–1265.
- [25] F. Azhari, C. C. Sennersten, C. A. Lindley, and E. Sellers, "Deep-learning implementations in mining applications: a compact critical review," *Artif. Intell. Rev.*, vol. 56, pp. 14367–14402, 2023.
- [26] Y. Fu and C. Aldrich, "Deep learning in mining and mineral-processing operations: a review," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 11920–11925, 2020.
- [27] M. R. U. Saputra, I. D. Bhaswara, B. I. Nasution, *et al.*, "Multi-modal deep-learning approaches to semantic segmentation of mining footprints with multispectral satellite imagery," *Remote Sens. Environ.*, vol. 306, art. 114114, 2024.
- [28] E. Vivoli, M. Bertini, and L. Capineri, "Deep-learning-based real-time detection of surface landmines using optical imaging," *Remote Sens.*, vol. 16, no. 4, art. 677, 2024.
- [29] L. He, Y. Zhou, L. Liu, W. Cao, and J. Ma, "Research on object detection and recognition in remote-sensing images based on YOLOv11," *Sci. Rep.*, vol. 15, art. 96314, 2025.
- [30] W. Lin, T. Li, and X. Li, "Deep-learning-based object detection for environmental monitoring using big data," *Front. Environ. Sci.*, vol. 13, art. 1566224, 2025.
- [31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.
- [32] Q. Wang, B. Wu, P. Zhu, *et al.*, "ECA-Net: efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11534–11542.
- [33] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: a light-weight context-guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.