buildings

MDPI

Article

Real-Time Safety-Critical Object Detection in Large Open Construction Sites Using a Scale-Gated Edge Detection Transformer

Lei Shen [1], Yanran Shi [2,*], Hao Lu [2,*], Zhanyun Gu [1], Dong Niu [1], Xin Yang [1], Ke Gao [2], Yuanping Liu [2] and Yanjie Wang [3,*]

[1] Water Conservancy Project Construction Center of Ningxia Autonomous Region, Yinchuan 750002, China; shl800520@163.com (L.S.); 13629558710@163.com (Z.G.); neotong981@163.com (D.N.); 18509579088@163.com (X.Y.)
[2] Materials & Structural Engineering Department, Nanjing Hydraulic Research Institute, Nanjing 210024, China; 18855576685@163.com (K.G.); m18395591055@163.com (Y.L.)
[3] Department of Mechanical Engineering, The University of Western Australia, Perth, WA 6009, Australia
* Correspondence: yrshi@nhri.cn (Y.S.); luhaoxiaobai@163.com (H.L.); wangyanjie@hebeu.edu.cn (Y.W.)

Abstract

Wide-area visual monitoring of construction sites is constrained by the reliable detection of safety-critical targets that appear small, low-resolution, and weakly textured under elevated or distant camera views. To address this problem, this study proposes Scale-Gated Edge Detection Transformer (SGE-DETR), a safety-oriented end-to-end detector for large open construction scenes. The framework incorporates scale-aware residual edge modulation to preserve weak contours and local structures, density-guided context-adaptive fusion to balance multi-level features according to contextual and edge-density responses, and spatial gated reparameterized feature refinement to suppress redundant background textures. Experiments were conducted on SODA and STWD using COCO-style scale-sensitive metrics and efficiency indicators. On SODA, SGE-DETR achieved AP50, APS, APM, and APL values of 0.8748, 0.2157, 0.4577, and 0.6013, respectively, with 32.5 GFLOPs, 14.5 M parameters, and 83.4 FPS. On STWD, it obtained the highest AP50, APS, APM, and APL among the compared methods, reaching 0.7936, 0.8132, 0.8594, and 0.9253, respectively. Ablation results further showed that the full model improved mAP50 and mAP50–95 over RT-DETR-r18 by 4.15 and 2.93 percentage points while reducing computational complexity. These results indicate that SGE-DETR improves safety-oriented small-object detection and multi-scale robustness while retaining a relatively low parameter count.

Keywords: construction safety monitoring; small object detection; large open construction sites; wide-area visual monitoring; transformer detector

Check for updates

Academic Editor: Jaewook Jeong

Received: 20 May 2026
Revised: 19 June 2026
Accepted: 24 June 2026
Published: 26 June 2026

Copyright: © 2026 by the authors. 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.

1. Introduction

Construction sites are typical open, dynamic, and high-risk work environments, characterized by complex conditions such as dense personnel movement, simultaneous operation of machinery and equipment, overlapping material storage, exposed edges and openings, and concurrent operations by multiple trades [1]. In such scenarios, the timely identification of workers, construction machinery, personal protective equipment (PPE), and other safety-related targets provides an important basis for construction safety supervision, risk warning, and refined on-site management [2]. With the development of smart construction sites and digital construction technologies, image- and video-based automatic monitoring has gradually become an important technical approach for improving the

efficiency of safety management. Compared with traditional manual inspection, visual monitoring systems can provide higher-frequency, wider-range, and more continuous perception of on-site conditions, thereby offering more timely data support for construction safety management [3].

Object detection plays a fundamental role in visual monitoring systems. It can not only provide spatial localization and semantic recognition results for safety-related entities in construction images, such as workers, machinery, PPE, warning signs, and hazardous-area elements [4,5], but also further support PPE compliance checking, human–machine spatial relationship analysis, danger-zone intrusion warning, and structured safety-event generation [6]. Therefore, object detection is not merely a visual perception task but a key enabling component in intelligent construction safety supervision and early risk warning systems.

However, safety supervision on actual construction sites is not merely a close-range target recognition problem. Large open construction scenes are usually characterized by wide spatial coverage, dispersed work areas, intermingled machinery and personnel, and complex background structures. To obtain a more complete overview of construction site conditions, on-site monitoring systems often rely on elevated fixed cameras, tower crane cameras, wide-angle cameras, or perimeter monitoring equipment to continuously observe large construction areas. Although this monitoring mode is advantageous for covering a wider workspace, it also causes many safety-related targets to appear as small-scale, low-resolution objects with blurred edges and insufficient texture information. For example, distant workers, safety helmets, reflective vests, small construction equipment, and local safety signs often occupy only a few pixels in monitoring images, and their appearance features are easily affected by illumination changes, occlusion, motion blur, and complex background interference, leading to missed detections and false detections [7].

As construction-site monitoring needs have gradually expanded from fixed-point perception to site observation with larger coverage, more flexible viewpoints, and greater spatial coverage capability, UAV-based inspection has also become a highly promising complementary approach for large-area construction-site monitoring. UAVs can provide flexible overhead views and rapidly collect visual information from areas that are difficult to access, dangerous, or inefficient for manual inspection, thereby further extending the spatial coverage of traditional monitoring systems [8]. In addition, UAV imagery can be combined with photogrammetric reconstruction techniques to generate point clouds or three-dimensional site records, thereby supporting construction progress monitoring and spatial comparison with planned digital models, such as 4D BIM [9]. These applications indicate that reliable object detection can serve as an important visual front end for UAV-assisted construction safety inspection. However, under aerial or high-view imaging conditions, workers, PPE-related targets, and local safety signs usually appear as small-scale and weak-texture objects. Therefore, robust small-object detection is an important prerequisite for the practical application of UAV-assisted construction safety monitoring.

In addition to small-target detection, target scale variation is another key challenge in the visual monitoring of large open construction scenes (Figure 1). Under the same monitoring perspective, people or equipment close to the camera may appear as medium- or even large-scale targets, whereas similar targets in distant areas may degrade into small-scale targets. Therefore, targets of the same category can exhibit significant scale differences due to variations in spatial location, shooting angle, and working distance. This requires the detection model to be sensitive to fine details of small targets while maintaining stable representation of medium- and large-scale target structures. If the model relies solely on single-scale features or lacks effective cross-scale information interaction, it is prone to missed detection of distant targets, inaccurate localization of medium-scale targets, and

insufficient robustness in scenarios with large scale variations [10]. Therefore, detectors for large open construction scenes need to preserve subtle edge and texture cues of distant small targets, maintain robust performance for medium- and large-scale objects, and suppress irrelevant high-frequency background patterns. Although general object detectors perform well on common benchmarks, they may still miss safety-critical small targets when local structural cues are weakened through multi-stage feature extraction or insufficiently adaptive feature fusion.

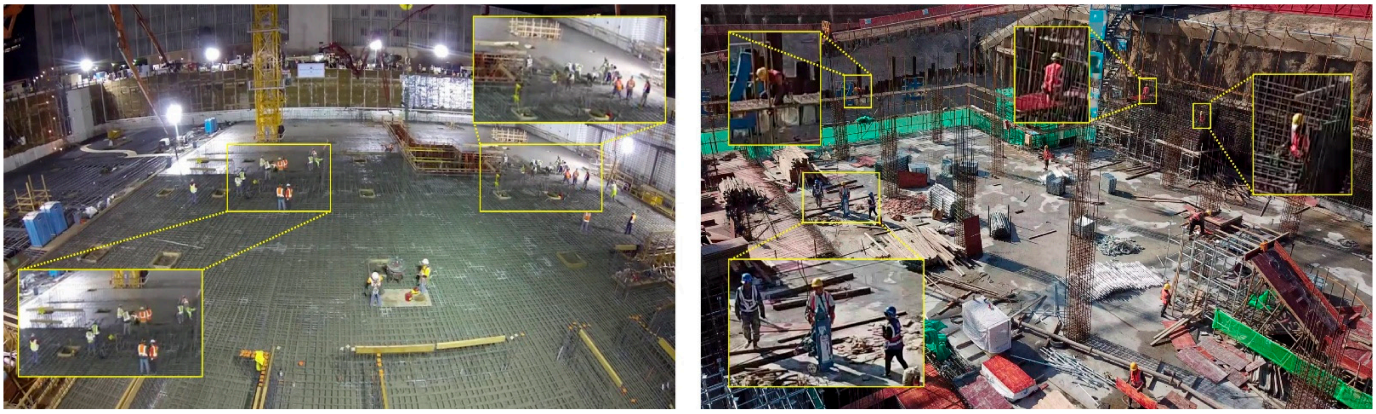


Figure 1. Large-scale open construction scene.

Deep-learning-based object detection provides a practical solution for construction-site visual monitoring because it can automatically learn discriminative visual representations from complex scene data and jointly perform target localization and category recognition [11]. Compared with manual inspection or rule-based image analysis, deep learning detectors are more suitable for continuous monitoring scenarios involving illumination variation, occlusion, dense personnel movement, cluttered materials, and large changes in object scale. More importantly, the detected objects can be directly used as structured inputs for safety compliance assessment, hazard identification, worker–machinery spatial relationship analysis, and real-time warning systems [12]. Therefore, improving the detection reliability of safety-critical targets is essential for transforming visual monitoring from passive recording into active safety supervision.

In recent years, deep-learning-based object detection methods have been widely applied to construction safety monitoring. Detection frameworks represented by Faster Region-based Convolutional Neural Network (Faster R-CNN) [13], Single Shot MultiBox Detector (SSD) [14], and You Only Look Once (YOLO) [15] have achieved promising performance in tasks such as personnel identification, mechanical equipment detection, helmet-wearing detection, and personal protective equipment compliance monitoring. In particular, single-stage detectors show strong application potential in real-time construction monitoring due to their high inference efficiency. For example, Long et al. [16] developed an SSD-based helmet-wearing detection system for substation workers, enabling feature extraction, target localization, and object classification. Nath et al. [17] proposed several improved YOLO-based models for detecting PPE-wearing compliance among construction workers. Li et al. [18] further proposed OAM-YOLO, which integrates object detection with human pose estimation and focuses on improving the recognition of small targets such as helmets and masks, thereby achieving more accurate PPE compliance assessment.

However, most existing studies still focus on benchmark testing in relatively controlled or small- to medium-scale scenarios. Their adaptability to large open construction scenes remains insufficiently verified, and they do not fully address the combined challenges of small targets, severe scale variation, and complex background interference in such scenarios [19–21].

Transformer-based object detectors provide a promising direction for addressing these challenges. Detection Transformer (DETR) reformulates object detection as an end-to-end set prediction problem and adopts a bipartite matching strategy with a Transformer encoder–decoder architecture [22]. Deformable DETR improves DETR by focusing on a small set of sampling points around reference points, thereby alleviating slow convergence and limited spatial resolution, especially for small objects [23]. DETR with Improved DeNoising Anchor Boxes (DINO) further enhances DETR-like models through denoising training, hybrid query selection, and fine box refinement [24]. Recently, Real-Time Detection Transformer (RT-DETR) introduced a real-time end-to-end detection framework that adopts an efficient hybrid encoder and uncertainty-minimal query selection, providing a practical basis for real-time detection without relying on non-maximum suppression [25]. However, when RT-DETR-like detectors are directly applied to large open construction scenes, distant safety-critical targets may still suffer from weak edge details, insufficient cross-scale interaction, and interference from high-response background structures.

To address these limitations, this study proposes a safety-oriented detection framework, termed Scale-Gated Edge Detection Transformer (SGE-DETR), for large open construction scenes. Under wide-area monitoring conditions, it remains challenging to reliably detect distant workers, personal protective equipment-related objects, and personnel operating around machinery, as these targets usually appear small, low-resolution, and weak in texture. The proposed framework integrates scale-aware residual edge modulation, density-guided adaptive feature fusion, and spatial gated reparameterized feature refinement to enhance fine-grained target representation, improve multi-scale feature interaction, and suppress complex construction background interference.

The main contributions of this study are summarized as follows:

1. A safety-oriented real-time detection framework, SGE-DETR, is proposed for large open construction scenes, focusing on distant workers, PPE-related objects, and personnel around machinery under wide-field monitoring conditions.
2. A Scale-aware Residual Edge Modulation Module (SREM) is designed to enhance weak edge, texture, and local structural cues of small safety-critical targets, thereby improving their representation under long-distance imaging and cluttered backgrounds.
3. A Density-guided Context Adaptive Fusion Module (DCAF) is developed to adaptively fuse multi-level features according to global contextual responses and edge-density cues, improving robustness under severe scale variation and multi-target coexistence.
4. A Spatial Gated Reparameterized C3 (SGRC3) module is introduced to selectively refine safety-related local responses while suppressing redundant background textures, achieving a favorable balance between detection accuracy and inference efficiency.
5. Experiments on the Site Object Detection and Analysis (SODA) and Small Target Worker Dataset (STWD) datasets show that SGE-DETR improves detection performance across small, medium, and large target scales while maintaining relatively low computational cost and high inference speed, supporting its potential for real-time construction safety monitoring.

The remainder of this paper is organized as follows. Section 2 reviews vision-based construction safety supervision and small object detection. Section 3 introduces the proposed SGE-DETR framework and its main components. Section 4 presents the experimental settings and comparative results. Section 5 discusses the findings, limitations, and future research directions. Section 6 concludes the paper.

2. Related Work

2.1. Vision-Based Construction Safety Supervision

With the rapid development of deep learning, data-driven methods have been widely applied to complex engineering [26–28] and visual analysis tasks [29], such as structural health monitoring [30–32], construction management [33], and pavement condition monitoring [34]. In construction safety supervision, vision-based detection has also become an important application of this trend, with PPE compliance detection being one of the most common tasks. Early deep-learning studies primarily focused on detecting whether workers wore safety helmets or hard hats in site images. Fang et al. [7] specifically addressed non-hardhat detection in far-field surveillance videos, which is closely related to the wide-field monitoring setting considered in this paper. Wu et al. [35] proposed a deep-learning-based hardhat detection system and constructed a benchmark dataset covering various on-site conditions. These studies demonstrated the feasibility of automatic PPE monitoring; however, they were mainly designed for helmet or hardhat recognition rather than broader multi-scale safety-object perception in large open construction scenes.

With the development of one-stage detectors, YOLO-based models have been widely used in construction PPE detection because of their real-time performance. Han and Zeng proposed a multi-scale feature-based safety-helmet-wearing detection method for construction sites [36]. Song proposed RSSE-YOLOv3 for multi-scale safety helmet detection, showing that multi-scale feature enhancement is useful for helmet recognition under varying object sizes [10]. Hayat and Morgado-Dias developed an automatic safety helmet detection system for construction safety, further confirming the practical value of deep learning for helmet compliance monitoring. More recently [37], Jiao et al. proposed a Unmanned Aerial Vehicle (UAV)-image-based helmet-wearing detection method using YOLOv8 and introduced matching and time-window strategies to reduce false positives and false negatives in UAV construction scenarios [38].

Worker and equipment detection has also been studied for construction safety monitoring. Fang et al. [39]. used a convolutional neural network (CNN) approach to detect workers and heavy equipment on construction sites, supporting safety and planning decisions. Wang [40] proposed a domain-adaptive Faster R-CNN to improve non-PPE violation detection across body-worn and general construction-site images. Li et al. [41] proposed a deep-learning-based method for inspecting the standardized use of PPE by workers, with a further focus on the spatiotemporal relationships among PPE, human posture, key body parts, and work behaviors, thereby improving safety supervision from the perspectives of administrative control and behavioral regulation. Park et al. [42] proposed YOLO based on small and overlapping object detection in construction sites (SOC-YOLO) for small and overlapping worker detection at construction sites, introducing attention, additional prediction heads, and Softpool- Spatial Pyramid Pooling-Fast (SPPF) to address poor visibility, overlap, and long-distance imaging. These works are highly relevant to the present study because they directly indicate that construction workers are difficult to detect when they are small, distant, occluded, or visually weak. However, most existing studies either focus on a single safety category, such as helmets, or design task-specific YOLO variants for workers [43–45]. The joint problem of safety-critical small targets, multi-scale objects, and structured construction background interference remains insufficiently addressed.

2.2. Small Object Detection

Small object detection remains a difficult problem in computer vision. Cheng et al. [46] emphasized that small objects suffer from poor visual appearance and noisy representation, and that large-scale benchmarks are needed to promote small-object detection research. Feng et al. [47] reviewed deep-learning-based small-object detection from several perspec-

tives, including increasing feature resolution, scale-aware training, contextual information, and data augmentation. Another recent review pointed out that small-object detection is affected by limited spatial information, low resolution, insufficient contrast, weak edge and texture cues, and a persistent accuracy gap between small and normal-sized objects [48]. These observations are consistent with the challenges of distant workers, helmets, and reflective vests in wide-field construction monitoring.

In recent years, small-object detection has developed rapidly in remote sensing imagery, UAV aerial imagery, and other related fields. Cao et al. [49] summarized representative methods in the VisDrone object detection challenge and pointed out that UAV aerial images are commonly affected by high-altitude viewpoints, small-scale targets, dense object distributions, occlusion, and complex urban backgrounds. Accordingly, related studies have mainly focused on multi-scale feature fusion, attention mechanisms, contextual modeling, and lightweight detection. Hua et al. [50] systematically reviewed deep-learning-based small-object detection methods in aerial images and identified low resolution, complex backgrounds, dense object distributions, and scale variation as the main bottlenecks of this task. They further categorized existing improvement strategies into feature enhancement, multi-scale fusion, attention modeling, context utilization, and detection head optimization. Another recent study [51] pointed out that distant small targets, complex backgrounds, occlusion interference, and resource-constrained deployment remain the main challenges for small-object detection in construction scenarios. This study also experimentally demonstrated the effectiveness of multi-scale fusion for detecting typical small targets, such as safety helmets, in complex construction environments.

The above studies indicate that multi-scale representation is a common strategy for detecting objects at different scales. Feature Pyramid Networks introduced a top-down architecture with lateral connections to construct semantic feature maps at multiple scales with relatively low extra cost [52–54]. Since then, feature pyramids, path aggregation, attention-based fusion, and adaptive feature selection have become standard components in modern detectors. However, construction scenes differ from generic object detection scenarios because high-frequency background structures such as scaffolds, steel bars, formwork, and material piles may resemble meaningful object boundaries [55]. As a result, simply adding high-resolution features or fusing multi-level features with fixed weights may amplify background noise while still failing to preserve weak PPE details [56–58].

3. Proposed Method

3.1. Overall Framework of the Proposed Detector

The proposed SGE-DETR adopts a typical end-to-end detection paradigm [25]. As shown in Figure 2, the overall architecture consists of a feature extraction network, a feature interaction encoder, and a Transformer decoder. To maintain real-time detection capability and the advantages of end-to-end prediction, SGE-DETR follows the basic feature interaction process of Attention-based Intra-scale Feature Interaction (AIFI) and CNN-based Cross-scale Feature Fusion (CCFF). Specifically, AIFI is used to enhance the global modeling capability of high-level semantic features, while CCFF facilitates information transmission and fusion among multi-scale features.

To address the susceptibility of safety-critical small targets in large open construction scenes to scale degradation, detail attenuation, and background interference, SGE-DETR reconstructs the feature representation and cross-scale fusion processes of the detector. Specifically, the feature extraction network is redesigned, and SREM is introduced to strengthen the edge contours, texture boundaries, and local structural information of distant safety-related targets. In the cross-scale feature interaction stage, DCAF adaptively adjusts the fusion contributions of features at different levels according to contextual

responses and edge density. In the fused feature refinement stage, SGRC3 is employed to enhance safety-related spatial responses and suppress redundant texture interference from complex construction backgrounds. Finally, the enhanced multi-scale features are fed into the Transformer decoder for query decoding, thereby completing the end-to-end prediction of target categories and bounding boxes.

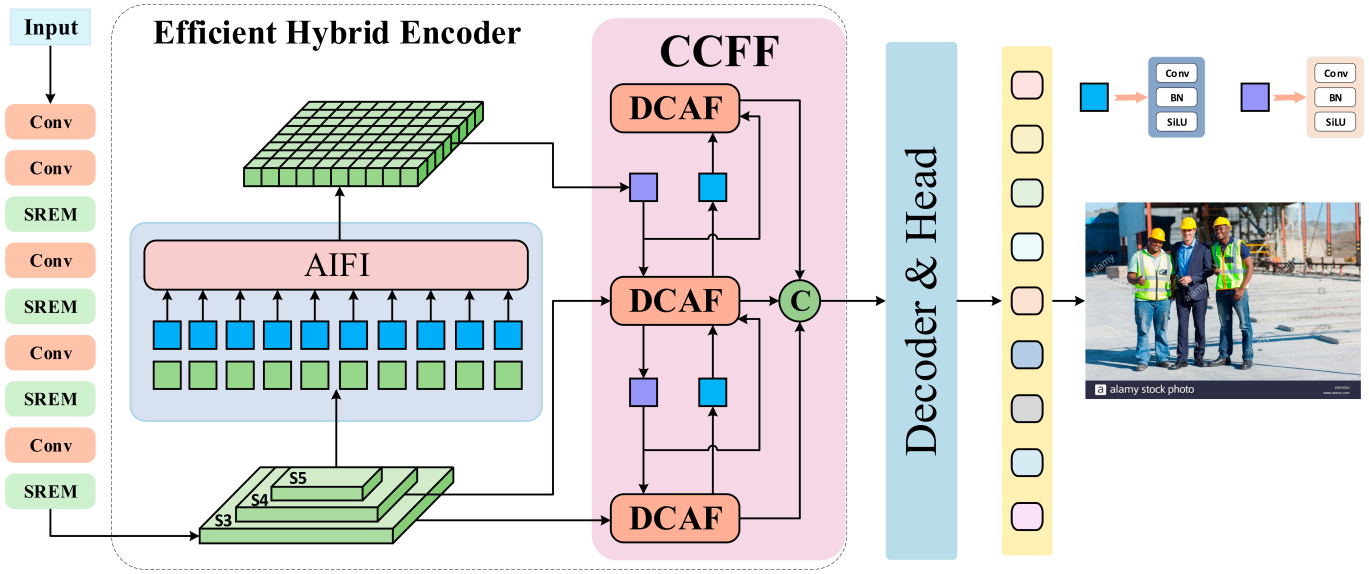


Figure 2. Architecture of the SGE-DETR network.

3.2. Scale-Aware Residual Edge Modulation Module

In large open construction scenarios, safety helmets, reflective vests, and workers often appear as distant small-scale targets. Compared with close-range targets, these targets contain limited texture details, and their contours are more easily disturbed by complex background structures. In construction images containing dense scaffolding, rebars, formwork, fences, and material-stacking areas, targets and backgrounds may exhibit similar colors or edge responses, making it difficult for detection models to accurately distinguish safety-related target regions from irrelevant backgrounds. Therefore, relying solely on conventional convolutional feature extraction may lead to edge information attenuation and insufficient local structural representation for small targets.

To alleviate this problem, an SREM module is designed, as shown in Figure 3. Starting from the visual characteristics of construction safety targets, this module integrates local texture preservation, multi-scale contextual modeling, and edge residual enhancement to improve the model's ability to perceive the contours of distant small targets. Given an input feature map X, the module first obtains a dimension-aligned feature X' through channel mapping. Here, X denotes the input feature map, and X' denotes the channel-aligned feature used for subsequent feature extraction. When the input and output channels are consistent, this mapping is implemented as an identity operation; otherwise, a 1×1 convolution is used for channel alignment. Subsequently, a local branch is introduced to extract neighborhood-level texture and structural information using 3×3 convolutions, thereby preserving details such as safety helmet edges, local human body contours, and reflective vest boundaries. The output of the local branch is denoted as F_l.

To address significant target scale variations in construction monitoring images, SREM further divides the features into multiple scale branches:

$$S = \{2, 4, 6, 8\} \tag{1}$$

where S denotes the set of spatial scales used in the multi-scale branches, and $s \in S$ represents a specific scale branch. The output feature of the s-th scale branch is denoted as F_s.

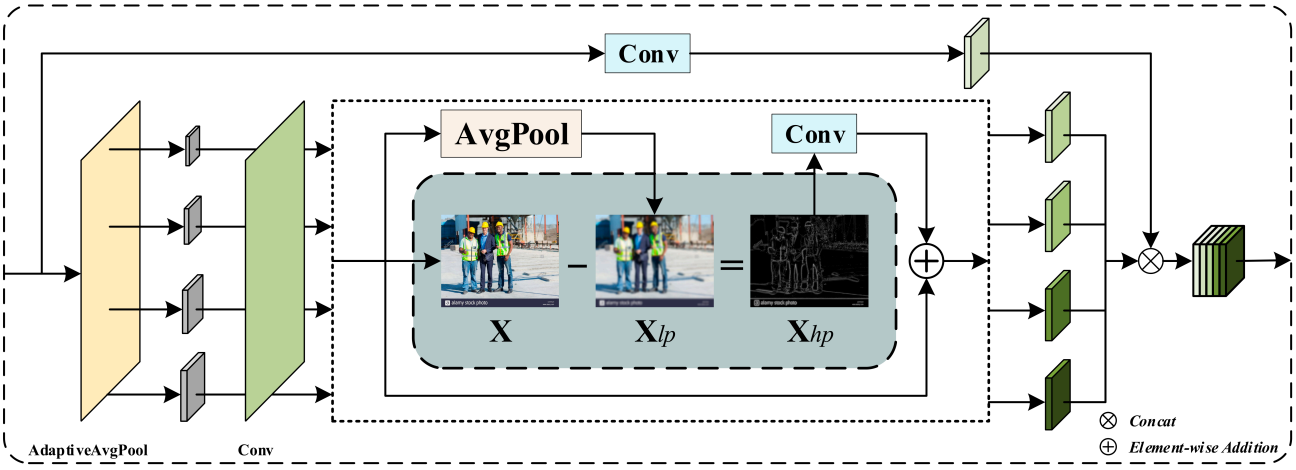


Figure 3. Structure of the SREM module.

For each scale branch, the input features are first compressed to the corresponding spatial scale through adaptive average pooling (AAP). Then, a 1×1 convolution is used for channel compression, followed by a depthwise convolution to extract local structural information at that scale. Thus, *Fs* represents the scale-specific feature after adaptive pooling, channel compression, and depthwise convolution. This process enables the model to capture the relationship between safety-related targets and construction backgrounds across different spatial ranges, while preserving local details and incorporating contextual information.

In each scale branch, a residual edge gating mechanism is further introduced. Specifically, smooth features are first obtained through 3×3 AAP, and high-frequency edge responses are then extracted by calculating the difference between the original features and the smoothed features:

$$E_s = F_s - AAP(F_s) \tag{2}$$

where *Fs* denotes the output feature of the s-th scale branch, $AAP(\cdot)$ represents the 3×3 average pooling operation, and *Es* denotes the edge residual information at this scale. $APP(Fs)$ is the smoothed component of *Fs*, while *Es* represents the high-frequency residual component, which contains local variations such as object contours and texture boundaries. Subsequently, the module generates edge gating weights through a convolution followed by a Sigmoid activation, and reinjects the effective edge information into the features in a residual manner:

$$\hat{F}_s = F_s + \sigma(Conv_{3\times3}(E_s)) \odot E_s \tag{3}$$

where $\hat{F}_s$ denotes the edge-enhanced feature of the *s*-th scale branch, $\mathrm{Conv}_{3\times3}(\cdot)$ is used to generate the edge-gating response, $\sigma(\cdot)$ denotes the Sigmoid activation function, and $\odot$ represents element-wise multiplication. The term $\sigma(\mathrm{Conv}_{3\times3}(Es))$ adaptively controls the contribution of the edge residual *Es*. This design enables the model to moderately enhance safety-related edge information, such as helmet contours, reflective vest boundaries, and worker silhouettes. Meanwhile, for irrelevant background patterns, such as rebar textures, scaffold edges, and material-stacking textures, the gating mechanism can reduce the risk of excessive amplification.

After edge enhancement, the features from different scale branches are restored to the spatial size of the input feature map through bilinear interpolation and concatenated with the output of the local branch. A 1×1 convolution is then used for feature fusion. When

the input and output channels are consistent, a learnable residual coefficient is further introduced to stabilize the module output:

$$Y = X + \gamma Conv_{1\times1}(Concat(F_l, \hat{F}_2, \hat{F}_4, \hat{F}_6, \hat{F}_8)) \quad (4)$$

where F_l denotes the output of the local branch and γ is a learnable parameter. In the implementation, γ is initialized to 0, which prevents the module from introducing excessive perturbations to the original features at the early training stage and helps maintain training stability.

Overall, SREM is designed to address weak edges, blurred contours, and strong background interference of small targets in large open construction scenes. Through multi-scale contextual branches and residual edge gating, SREM enhances the locally discriminative information of safety-related targets. Rather than simply increasing network depth, this module compensates for the loss of contour details that commonly occurs when distant safety-monitoring targets are captured under wide-field surveillance conditions.

3.3. Density-Guided Context Adaptive Fusion Module

The target scale distribution in construction sites is highly imbalanced. PPE-related targets, such as safety helmets, gloves, and local regions of reflective vests, are usually small in scale, whereas workers, construction machinery, vehicles, scaffolding, and material-stacking areas occupy larger spatial regions. In the same monitoring image, small-scale safety-related targets and large-scale background objects often appear simultaneously. If fixed weighting, simple summation, or direct concatenation is adopted for feature fusion, the model may fail to adaptively adjust the contribution of different features according to the target distribution and background complexity of the current construction scene. This may lead to two problems: first, small-target details can be weakened during feature fusion; second, complex background textures may be incorrectly treated as effective target information.

To address this issue, a DCAF module is designed, as shown in Figure 4. This module is used to fuse multiple feature inputs with the same spatial resolution and dynamically estimate the importance of each feature branch according to its global response and edge density. Unlike traditional Feature Pyramid Network (FPN) or Path Aggregation Network (PAN)-based fusion methods, DCAF does not assume that different features make fixed contributions across all construction scenes. Instead, it adaptively adjusts the fusion weights according to the content of the current image.

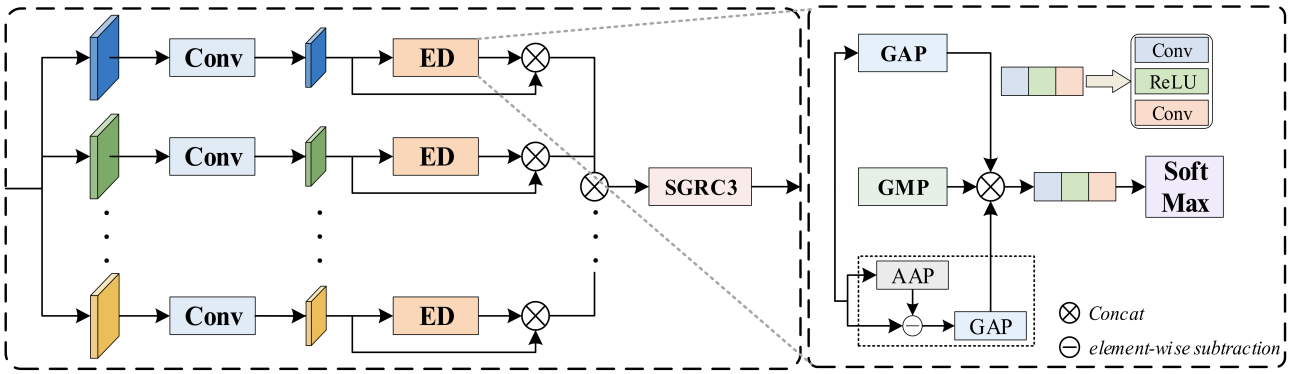


Figure 4. Structure of the DCAF module.

Let the multi-branch input features be denoted as $\{X_i\}_{i=1}^{n}$. Here, n denotes the number of input feature branches, and X_i represents the input feature of the i-th branch. In the implementation, each input feature is first mapped to a unified channel dimension c_2

through a 1×1 convolution. If the spatial size of a feature branch differs from the target size, nearest-neighbor interpolation is used to align it with the spatial size of the first input feature. After obtaining the aligned feature F_i, the module extracts its global average response, global maximum response, and edge density information. Here, F_i denotes the aligned feature of the i-th branch after channel mapping and spatial alignment. The edge density is calculated based on the difference between the feature and its average-pooled result:

$$D_i = \mathrm{GAP}(\mathrm{Mean}_c(|F_i - AAP(F_i)|)) \tag{5}$$

where $GAP(\cdot)$ denotes global average pooling, $Mean_C(\cdot)$ denotes the mean operation along the channel dimension, and $AAP(\cdot)$ denotes adaptive average pooling. The edge density term reflects the strength of local structural and boundary variations in the current feature. In construction scenes, when an image contains many distant workers, safety helmets, or reflective vest boundaries, edge density can provide valuable guidance for feature fusion. When the background contains complex textures, this information also helps determine whether different features include excessive interference responses.

Subsequently, the global average response, global maximum response, and edge density are concatenated and fed into a lightweight weight estimation branch to obtain the fusion score of each feature branch. The scores of all branches are normalized by Softmax to generate the fusion weights:

$$w_i = \mathrm{Softmax}_i(\psi_i(\mathrm{Concat}(\mathrm{GAP}(F_i), \mathrm{GMP}(F_i), D_i))) \tag{6}$$

where $GMP(\cdot)$ denotes global max pooling, and ψ_i denotes the lightweight weight estimation function corresponding to the i-th input branch. Concat$(\cdot)$ denotes feature concatenation, and w_i represents the normalized fusion weight of the i-th branch. The subscript iii in $\mathrm{Softmax}_i(\cdot)$ indicates that the normalization is performed across all input branches. Finally, the multi-branch features are weighted and fused, followed by further refinement using SGRC3:

$$Y = \mathrm{SGRC3}\left(\sum_{i=1}^{n} w_i F_i\right) \tag{7}$$

where Y denotes the output feature of the DCAF module. In Equation (7), $w_i F_i$ represents the weighted feature of the i-th branch, and $\sum_{i=1}^{n} w_i F_i$ denotes the adaptive fusion result of all aligned input features.

Through this design, DCAF can dynamically adjust the contribution of different features according to the target scale, target density, and edge complexity of the construction scene. When an image contains many distant small PPE-related targets, the module can assign greater importance to features containing edge and detail information. When large equipment, material-stacking areas, or complex background structures are present, the module can incorporate global contextual responses to reduce the influence of irrelevant textures on the detection results. Therefore, DCAF improves the stability of feature fusion under conditions where multi-scale targets coexist with complex background interference.

3.4. Spatial Gated Reparameterized C3 Module

In large open construction scenes, safety-critical small targets are usually embedded in complex backgrounds. In addition to key targets such as workers, safety helmets, and reflective vests, construction images often contain a large number of structured background elements, including scaffolding, rebars, formwork, fences, machinery, and material-stacking areas. These background regions usually exhibit distinct edges, textures, and repetitive linear patterns, which can produce strong responses during feature extraction. When the model fails to effectively distinguish safety-related target features from background

texture features, two problems may occur. First, complex backgrounds may be incorrectly identified as safety-related targets, resulting in increased false detections. Second, the effective responses of small-scale PPE targets may be disturbed by background noise, increasing the risk of missed detections.

To enhance the feature selection capability of the model for construction safety targets, an SGRC3 module is designed, as shown in Figure 5. This module is constructed based on the RepC3 structure but replaces its original feature transformation units with SGRBlocks, namely Spatial Gated Reparameterized Blocks. In this way, the model is equipped with adaptive feature selection capability during local feature extraction. Unlike standard convolutions, which uniformly transform responses across all spatial locations and channels, SGRC3 uses the collaborative interaction between a content branch and a gating branch. This enables the model to preserve effective structural information related to safety targets while suppressing irrelevant texture responses from construction backgrounds.

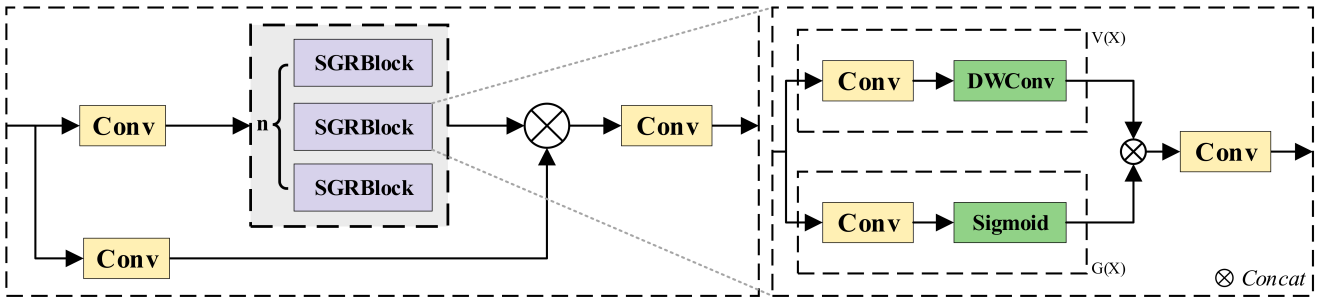


Figure 5. Structure of the SGRC3 module.

Specifically, given an input feature X, the gConvBlock first constructs two parallel branches. Here, X denotes the input feature map of the gConvBlock. The content branch $W_v(\cdot)$ consists of a 1×1 convolution and a depthwise convolution, where the 1×1 convolution is used for channel-wise information adjustment and the depthwise convolution is used to extract local spatial structures. In the implementation, reflection padding is adopted in the depthwise convolution to alleviate the influence of feature discontinuity near image boundaries on detection results. The gating branch $W_g(\cdot)$ consists of a 1×1 convolution and a gating activation function. By default, a Sigmoid function is used to generate feature selection weights. This process can be simplified as

$$Y = \mathrm{Conv}_{1\times1}\big(W_v(X) \odot W_g(X)\big) \tag{8}$$

where $\odot$ denotes element-wise multiplication. In Equation (8), *Wv(X)* represents the content feature extracted by the content branch, and *Wg(X)* represents the gating weight generated by the gating branch. The output Y denotes the refined feature of the gConvBlock. $\mathrm{Conv}_{1\times1}(\cdot)$ is used to further integrate the gated feature and adjust the channel dimension. This formulation reflects the core idea of the module: the content branch extracts local structural features, while the gating branch determines which responses should be retained. For construction images, effective responses such as safety helmet edges, reflective vest regions, and worker contours can be assigned higher weights, whereas interference responses caused by rebar textures, scaffold lines, and material-stacking patterns can be suppressed to a certain extent.

In SGRC3, the module inherits the residual feature reuse form of RepC3 and replaces the internal repeated feature transformation units with multiple SGRBlocks. Let the hidden channel dimension be $c' = c_2 \cdot e$. Here, c_2 denotes the output channel dimension, and e denotes the expansion ratio used to determine the hidden channel width. In the implementation, n consecutive SGRBlocks are stacked to refine the intermediate features. Since RepC3

itself provides effective feature reuse and high inference efficiency, SGRC3 introduces more flexible local response modulation while maintaining a lightweight structure.

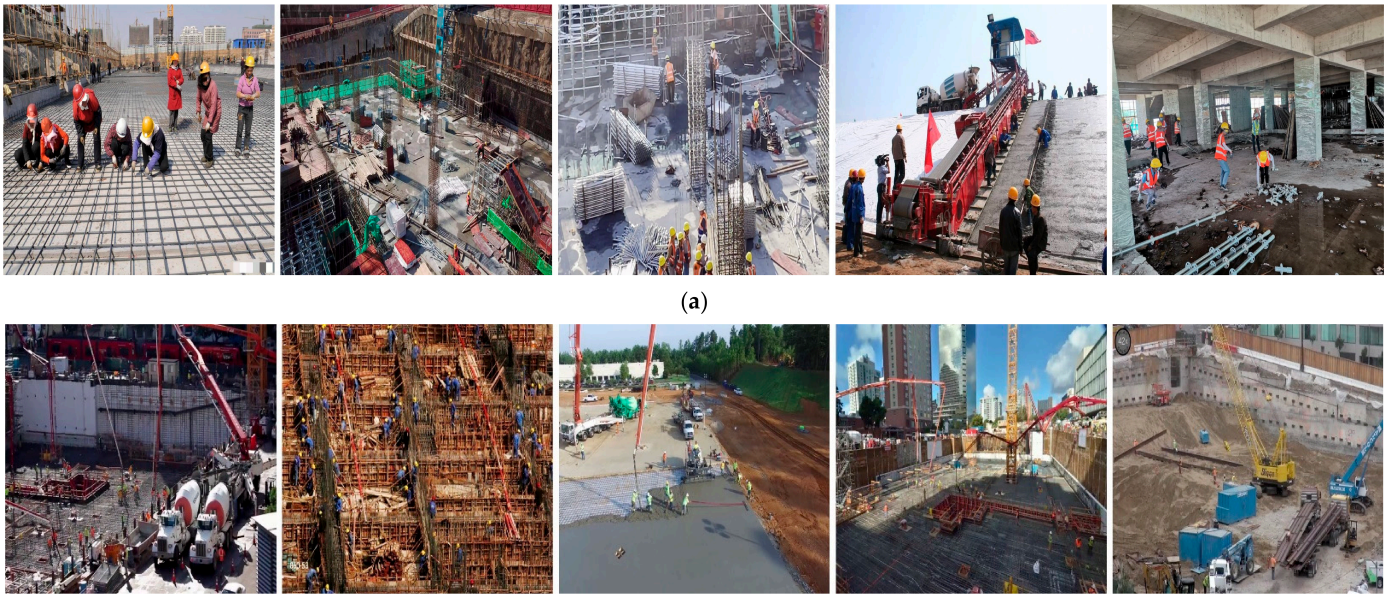
The role of this module in the construction safety monitoring framework is mainly reflected in two aspects. First, it enhances the selective representation of safety-related local structures, making distant safety helmets, reflective vest boundaries, and worker contours more prominent in complex backgrounds. Second, it suppresses interference caused by repetitive textures and high-frequency background structures commonly found on construction sites, reducing the model's incorrect attention to non-safety-related regions. Therefore, SGRC3 is more suitable as a feature refinement unit for complex construction scenes, improving detection stability under conditions of background complexity, frequent occlusion, and weak responses of small targets.

4. Experimental Results and Analysis

4.1. Datasets

(1) Site Object Detection and Analysis

In this study, the Site Object Detection and Analysis (SODA) dataset [59] is used for the main ablation experiments and performance evaluation. SODA is a large-scale real-world dataset constructed for object detection in open construction sites (Figure 6a). It supports intelligent perception research for multiple categories of construction-site objects, including workers, personal protective equipment, machinery, and materials.



(**a**)

(**b**)

Figure 6. Example images from the datasets. (**a**) SODA dataset; (**b**) STWD.

The dataset contains 19,846 images and 286,201 annotated instances, covering four major categories and 15 common construction object classes. The images are collected from various sources, including UAVs, handheld cameras, surveillance videos, and tower-crane-hook visualization devices. They cover different construction stages, shooting perspectives, weather conditions, and complex backgrounds. SODA is characterized by open scenarios, diverse object categories, significant scale variation, and strong background interference, making it suitable for reflecting the practical challenges of multi-scale object detection in large construction scenes.

In this paper, experiments are conducted according to the official data split, where 17,861 images are used for training and 1985 images are used for testing. This setting is

adopted to evaluate the detection performance and generalization ability of the proposed method in open construction scenarios.

(2) Small Target Worker Dataset

To further evaluate the small-target detection capability of the proposed framework in large construction scenes, we compare it with advanced small-object detectors on the Small Target Worker Dataset (STWD) [60]. STWD is a real-world dataset constructed for small worker detection in large construction sites. It aims to address the limitations of existing construction datasets in detecting densely distributed and distant small-scale workers.

The dataset contains 1705 images collected from surveillance videos of 25 different construction sites, covering various construction stages, weather conditions, and scene backgrounds (Figure 6b). A total of 48,553 worker instances are annotated in the dataset. Among them, approximately 73.6% have an area smaller than 32 × 32 pixels, while only about 1.4% have an area larger than 96 × 96 pixels. On average, each image contains approximately 28 small worker targets. Therefore, STWD is characterized by small target scales, dense distributions, and obvious overlap and occlusion, making it well suited for evaluating the challenges of small-object detection in large construction scenes.

In this paper, experiments are conducted according to the official dataset split, where 1210 images are used for training, 214 images are used for validation, and another 218 images from new scenes are used for testing. This setting is used to evaluate the model's ability to detect small workers in complex construction scenarios.

4.2. Experimental Setup

Unless otherwise specified, the training and testing experiments for all models in this study were conducted on a server equipped with an Intel Core i9-13900K CPU, an NVIDIA RTX 4090 GPU with 24 GB memory, and 32 GB RAM. The software environment included Python 3.8, PyTorch 2.2.2, and CUDA 11.8. During training, each model was trained for 200 epochs with a batch size of 16, and the input image size was set to 640 × 640. To improve training stability in the early stage, the number of warm-up epochs was set to 4. The initial learning rate and weight decay were both set to 0.0001, and AdamW was adopted as the optimizer.

4.3. Evaluation Metrics

To comprehensively evaluate the detection performance of the proposed method in small-object and multi-scale object scenarios, this study adopts the Common Objects in Context (COCO) object detection evaluation protocol and focuses on metrics including APs, APm, APl, and AP50. Specifically, APs denotes the detection accuracy for small-scale objects with an area smaller than 32 × 32 pixels, APm denotes the detection accuracy for medium-scale objects with an area between 32 × 32 and 96 × 96 pixels, and APl denotes the detection accuracy for large-scale objects with an area larger than 96 × 96 pixels. AP50 represents the average precision at an IoU threshold of 0.5 and is used to evaluate the basic object recognition and detection capability of the model.

The calculation of AP is given as follows:

$$AP = \int_0^1 P(R)dR \tag{9}$$

where P denotes precision and R denotes recall, with the confidence threshold ranging from 0 to 1. In addition, precision and recall are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

In addition to accuracy metrics, computational complexity measured by floating-point operations (FLOPs) and inference speed measured in frames per second (FPS) are also introduced to comprehensively evaluate the efficiency and practical applicability of the proposed model.

4.4. Comparison with Existing Detectors

(1) Quantitative and Qualitative Results (SODA):

As shown in Table 1, the proposed method achieves the best overall performance on the SODA dataset, indicating its strong adaptability to complex open-construction-site scenes. Compared with general one-stage detectors, our method shows a clear advantage across all scale-related metrics, especially in APS and APM. This demonstrates that the proposed framework can better preserve discriminative features of small and medium objects, which are usually weakened by long-distance imaging, background clutter, and multi-stage downsampling. Although recent YOLO variants maintain high inference speed, their limited improvement on small-scale objects suggests that conventional real-time detectors still struggle to capture weak and fine-grained construction-site targets.

Table 1. Performance comparison of small object detection in the SODA dataset.

Method	AP_{50}	AP_S	AP_M	AP_L	FLOPs/G	Params/M	FPS
one-stage							
YOLOv8s [61]	0.6653	0.0786	0.2764	0.4185	28.7	11.2	74.3
YOLOv10s [62]	0.6843	0.0874	0.2926	0.4358	24.5	8.0	79.5
YOLOv11s [63]	0.6952	0.0928	0.3047	0.4473	21.5	9.4	81.2
YOLOv12s [64]	0.7236	0.1079	0.3278	0.4726	<u>21.4</u>	9.3	82.0
TPH-YOLOv5 [65]	0.7269	0.1135	0.3342	0.4819	36.8	7.4	67.6
Small object detector							
SOD-YOLO [66]	0.7542	0.1268	0.3496	0.4973	**12.1**	**2.5**	<u>82.7</u>
PRNet [67]	0.7816	0.1384	0.3679	0.5168	44.9	7.7	65.8
STW-YOLO [60]	0.7938	0.1452	0.3795	0.5294	49.7	<u>5.6</u>	62.4
End-to-End							
DETR [22]	0.7024	0.0695	0.3027	0.4628	187.0	43.3	18.6
Deformable DETR [23]	0.7658	0.1129	0.3586	0.5115	267.8	42.3	15.4
DINO [24]	0.8127	0.1376	0.3864	0.5469	273.5	50.4	13.8
RT-DETR-r18 [25]	0.8333	0.1482	<u>0.3985</u>	0.5621	57.0	19.9	54.6
RT-DETR-r50 [25]	<u>0.8394</u>	<u>0.1835</u>	0.4063	<u>0.5687</u>	134.8	42.8	31.7
SGE-DETR	**0.8748**	**0.2157**	**0.4577**	**0.6013**	32.5	14.5	**83.4**

Note: All results were obtained by retraining each model and evaluating it on the corresponding test set. References [60,65–67] refer to improved variants proposed in the literature, rather than the original baseline models. The best results are in bold, and the second-best results are indicated by underlining.

Compared with specialized small-object detectors, our method also obtains higher AP50 and scale-wise AP values. This indicates that the proposed design not only improves small-object perception but also maintains robust detection for medium and large targets. In particular, the improvement in APS confirms its effectiveness in detecting distant workers, safety-related objects, and other small construction-site targets. Compared with DETR-based end-to-end detectors, our method achieves better detection accuracy with substantially lower computational cost than most transformer-based baselines. Although

RT-DETR-r50 provides competitive performance, it requires much higher FLOPs and has lower FPS. In contrast, our method achieves the highest AP50 while maintaining only 32.5 G FLOPs and 83.4 FPS, showing a favorable balance among detection accuracy, model complexity, and inference speed.

Overall, the SODA results demonstrate that the proposed method is not merely accurate under standard evaluation but is particularly effective for small and scale-varying targets in complex construction scenes. The gains in APS and APM suggest that the model can reduce missed detections of distant or weak-texture objects, while the high FPS indicates its practical potential for real-time safety monitoring.

(2) Quantitative and Qualitative Results (STWD):

Table 2 presents the comparison results on the STWD, which focuses on small-target worker detection in large construction scenes. The proposed method achieves the best AP50, APS, APM, and APL among all compared methods, indicating superior detection capability under dense distribution, long-distance imaging, and occlusion. Compared with one-stage detectors, the improvement is particularly evident in APS, suggesting that the proposed method can more effectively extract and retain small worker features that are easily confused with background textures in large construction sites.

Table 2. Performance comparison of small object detection in the STWD.

Method	AP_{50}	AP_S	AP_M	AP_L	FLOPs/G	Params/M	FPS
one-stage							
YOLOv8s [61]	0.6162	0.6203	0.7861	0.8376	28.7	11.2	74.3
YOLOv10s [62]	0.6473	0.6537	0.8145	0.8613	24.5	8.0	79.5
YOLOv11s [63]	0.6784	0.6804	0.8451	<u>0.9055</u>	21.5	9.4	81.2
YOLOv12s [64]	0.6937	0.7086	0.8315	0.8524	<u>21.4</u>	9.3	82.0
TPH-YOLOv5 [65]	0.6975	0.6962	0.8387	0.9025	36.8	7.4	67.6
Small object detector							
SOD-YOLO [66]	0.6593	0.6687	0.7761	0.8394	**12.1**	**2.5**	<u>82.7</u>
PRNet [67]	0.6861	0.7192	0.8196	0.8374	44.9	7.7	65.8
STW-YOLO [60]	<u>0.7712</u>	<u>0.7945</u>	<u>0.8506</u>	0.9012	49.7	<u>5.6</u>	62.4
End-to-End							
DETR [22]	0.5536	0.5742	0.7136	0.6845	187.0	43.3	18.6
Deformable DETR [23]	0.5672	0.5783	0.7342	0.6934	267.8	42.3	15.4
DINO [24]	0.5731	0.5803	0.7261	0.6989	273.5	50.4	13.8
RT-DETR-r18 [25]	0.5892	0.6231	0.7451	0.7112	57.0	19.9	54.6
RT-DETR-r50 [25]	0.6152	0.6295	0.7871	0.7213	134.8	42.8	31.7
SGE-DETR	**0.7936**	**0.8132**	**0.8594**	**0.9253**	32.5	14.5	**83.4**

Note: All results were obtained by retraining each model and evaluating it on the corresponding test set. References [60,65–67] refer to improved variants proposed in the literature, rather than the original baseline models. The best results are in bold, and the second-best results are indicated by underlining.

For specialized small-object detection methods, STW-YOLO obtains competitive results on STWD, especially for small workers. However, our method further improves AP50 and APS while achieving a much higher inference speed. This result indicates that the proposed framework provides stronger feature representation for small-scale workers while maintaining a relatively low parameter count. Although some small-object detectors have advantages in terms of parameter count, the proposed method achieves 83.4 FPS with a moderate computational cost, suggesting its potential for further validation and optimization in construction-site monitoring applications.

Compared with DETR-based methods, the proposed method shows a more pronounced advantage on STWD. End-to-end detectors perform relatively weakly on this dataset, which may be related to the loss of fine-grained features in low-resolution feature maps and the difficulty of modeling dense small workers. Our method effectively alleviates this limitation and achieves consistent improvements across different object scales. Notably, the high APS and APL indicate that the model does not improve small-target detection at the expense of normal-scale worker detection but maintains stable scale robustness.

In summary, the STWD results further verify the effectiveness and generalization of the proposed method for large-scale construction scenes. The strong performance in APS indicates fewer missed detections of distant small-scale workers, while the high FPS suggests that the method has preliminary potential for edge deployment.

4.5. Ablation Study

To verify the specific contribution of each component in the proposed method, a series of incremental ablation experiments were conducted based on the RT-DETR-R18 baseline and evaluated on the SODA dataset. The detailed results are presented in Table 3.

Table 3. Quantitative results of different components in the baseline on the SODA dataset. The baseline is RT-DETR-r18.

	SREM	DCAF	SGRC3	P	R	mAP_{50}	$mAP_{50–95}$	FLOPs/G	Params/M
Baseline				0.8456	0.8109	0.8333	0.5152	57.0	19.9
	√			0.8635	0.8287	0.8521	0.5287	43.4	16.6
		√		0.8682	0.8325	0.8548	0.5305	57.5	20.1
			√	0.8601	0.8243	0.8465	0.5261	45.0	17.4
	√	√		0.8793	0.8407	0.8662	0.5378	43.9	16.8
	√		√	0.8758	0.8352	0.8627	0.5329	32.0	14.3
		√	√	0.8729	0.8361	0.8601	0.5319	45.5	17.6
	√	√	√	0.8876	0.8456	0.8748	0.5445	32.5	14.5

The baseline model achieves a precision of 0.8456, a recall of 0.8109, an mAP50 of 0.8333, and an mAP50–95 of 0.5152, with 57.0 GFLOPs and 19.9 M parameters. When introduced individually, all three modules improve detection performance compared with the baseline. SREM increases the mAP50 and mAP50–95 to 0.8521 and 0.5287, respectively, while reducing the computational cost to 43.4 GFLOPs and the parameter count to 16.6 M, indicating its ability to preserve fine-grained cues for small targets. DCAF achieves an mAP50 of 0.8548 and an mAP50–95 of 0.5305, suggesting improved multi-level feature interaction under scale variation. SGRC3 increases the mAP50 and mAP50–95 to 0.8465 and 0.5261, respectively, while reducing the computational cost and parameter count to 45.0 GFLOPs and 17.4 M, respectively, indicating its capability for lightweight feature refinement.

The pairwise ablation results further indicate that the proposed modules do not introduce substantial redundancy or negative interactions. The combinations of SREM and DCAF, SREM and SGRC3, and DCAF and SGRC3 all outperform both the baseline and their corresponding single-module variants. In particular, the combination of SREM and DCAF achieves an mAP50 of 0.8662 and an mAP50–95 of 0.5378, indicating the complementarity between edge-aware representation and adaptive feature fusion. The combination of SREM and SGRC3 achieves the lowest computational cost, with 32.0 GFLOPs and 14.3 M parameters, indicating that edge enhancement and gated reparameterized refinement jointly improve efficiency. The combination of DCAF and SGRC3 also improves accuracy compared with the baseline, suggesting that adaptive fusion and spatial gated refinement are compatible.

The full model achieves the best overall performance, with a precision of 0.8876, a recall of 0.8456, an mAP50 of 0.8748, and an mAP50–95 of 0.5445. Compared with the baseline, the mAP50 and mAP50–95 are improved by 4.15 and 2.93 percentage points, respectively, while the computational cost and parameter count are reduced to 32.5 GFLOPs and 14.5 M, respectively. Since the full model outperforms all pairwise combinations, these results indicate that SREM, DCAF, and SGRC3 provide complementary benefits rather than redundant effects, enabling a better balance between detection accuracy and computational efficiency for safety monitoring in large construction scenes.

4.6. Performance and Error Analysis of Safety-Critical Objects

To further verify the perception capability of the proposed method for safety-critical targets, we analyzed five representative object categories in the SODA dataset: "person", "helmet", "vest", "hook", and "cutter".

As shown in Table 4, different detectors exhibit their respective advantages and limitations. YOLOv8s provides basic detection capability for workers, but its performance is limited in detecting small PPE and equipment-related targets, reflecting insufficient sensitivity to subtle local details. TPH-YOLOv5 improves small-object detection performance to some extent, but it is still affected by complex construction backgrounds and small machinery-related objects. Among the YOLO-based methods, STW-YOLO achieves better performance, supporting the effectiveness of task-oriented designs for small-scale worker detection; however, there remains room for improvement in the PPE and hazardous equipment categories. DETR shows relatively balanced but generally lower performance, suggesting that the original Transformer-based detector may lose fine-grained feature information for small safety-critical targets. RT-DETR-r18 outperforms DETR and provides a strong end-to-end baseline, but further improvements are still needed for detecting small PPE and local equipment targets.

Table 4. Comparison of AP for representative safety-critical object on the SODA dataset.

Method	$AP50_{person}$	$AP50_{helemt}$	$AP50_{vest}$	$AP50_{hook}$	$AP50_{cutter}$	AP50
YOLOv8s [61]	0.8112	0.6628	0.6945	0.6430	0.6155	0.6854
TPH-YOLOv5 [65]	0.8684	0.7832	0.7951	0.7056	0.6592	0.7623
STW-YOLO [60]	0.8815	0.8142	0.8388	0.8471	0.8279	0.8419
DETR [22]	0.8246	0.7015	0.7441	0.7213	0.6925	0.7368
RT-DETR-r18 [25]	0.8906	0.7799	0.8526	0.9396	0.8421	0.8610
SGE-DETR	0.9205	0.8427	0.8918	0.9624	0.9041	0.9043

In contrast, SGE-DETR achieves the best performance across all selected safety-critical categories, including person, helmet, vest, hook, and cutter. The improvements are particularly evident for helmet, vest, and cutter, as these targets are usually small in size, have weak texture features, and are easily confused with construction materials or mechanical components. This indicates that the proposed edge-aware representation, density-guided fusion, and spatial gated refinement mechanisms can better preserve key safety-related features while suppressing interference from complex construction backgrounds.

As shown in Table 5, the error analysis further supports the above findings. YOLOv8s shows relatively pronounced localization errors, background false positives, and missed detections, indicating that conventional one-stage detectors are still prone to localization deviations and target omissions when facing weak small-target features and complex construction backgrounds. DETR also exhibits evident localization and missed-detection issues, suggesting that the original Transformer-based detector remains insufficient in preserving fine-grained spatial information for small safety-critical targets. TPH-YOLOv5 and STW-YOLO alleviate these problems to some extent, reflecting the effectiveness of

small-object-oriented optimization designs. However, background false positives and missed detections remain difficult to fully avoid in construction scenes with complex background structures, small target scales, or severe partial occlusion.

Table 5. Error analysis on the SODA dataset.

Method	Cls	Loc	Both	Dupe	Bkg	Miss	FalsePos	FalseNeg
YOLOv8s [61]	1.42	7.85	1.36	0.38	6.74	8.96	12.48	15.37
TPH-YOLOv5 [65]	0.82	6.48	0.91	0.27	4.86	5.73	9.62	10.84
STW-YOLO [60]	0.63	4.18	0.58	0.19	3.58	2.88	7.42	7.10
DETR [22]	1.07	6.92	1.05	0.24	5.21	7.34	10.31	12.86
RT-DETR-r18 [25]	0.50	4.76	0.61	0.15	3.02	3.27	6.98	7.92
SGE-DETR	0.36	3.84	0.43	0.10	2.21	2.34	5.41	5.96

Note: Cls, Loc, Both, Dupe, Bkg, Miss, FalsePos, and FalseNeg denote classification error, localization error, joint classification-localization error, duplicate detection, background error, missed detection, overall false-positive error, and overall false-negative error, respectively.

In contrast, SGE-DETR exhibits lower error levels in terms of localization errors, background false positives, missed detections, overall false positives, and overall false negatives. This improvement can be mainly attributed to the complementary effects of the three proposed modules. SREM preserves weak edge and local structural information of small safety-critical targets, helping to reduce missed detections caused by scale degradation and blurred boundaries. DCAF adaptively balances multi-scale features based on contextual responses and edge-density information, thereby improving feature fusion under severe scale variations and dense target distributions. SGRC3 suppresses interference from repetitive background textures, such as scaffolding, rebars, material stacks, and mechanical structures, through its spatial gated refinement mechanism, thereby reducing background-induced false positives. These results indicate that SGE-DETR does not improve performance merely by increasing detection responses; instead, it reduces the risk of missing key safety targets while enhancing target–background discriminability. Nevertheless, localization error remains the main residual error source, indicating that accurate localization of extremely small targets, weak-boundary targets, and partially occluded targets still requires further improvement.

5. Discussion

5.1. Discussion in the Context of Small Object Detection

Existing studies on small-object detection mainly focus on scenarios such as remote sensing satellite images and UAV aerial images, while relatively limited attention has been paid to the detection of small-scale safety inspection targets in large open construction scenes. Different from general small objects in remote sensing or aerial imagery, distant workers, safety helmets, reflective vests, and personnel around machinery in large construction sites are not only small in scale but also directly related to safety supervision. Therefore, the proposed SGE-DETR aims to improve the detection capability for safety-critical small targets in large open construction scenes. Experimental results on the SODA and STWD datasets show that SGE-DETR enhances the model's perception of distant workers, small safety-related targets, and weak-texture targets in complex backgrounds, and demonstrates better stability in small-object and multi-scale object detection. These results indicate that the proposed method provides a feasible visual perception approach for recognizing safety-critical small targets in complex construction scenes.

This finding also has practical value for future intelligent inspection and safety warning applications on construction sites. In real engineering scenarios, elevated fixed cameras, tower crane cameras, perimeter wide-angle cameras, and UAV inspection devices are commonly used to acquire images of large construction areas. Such wide-area views can

cover more work zones and potential risk areas, but they also cause workers and PPE-related targets to appear as small-scale, low-resolution, and weak-texture objects in images, thereby increasing the risk of missed detections. Therefore, detection models with strong small-object detection capability and multi-scale adaptability can serve as fundamental perception modules in wide-area construction monitoring and UAV inspection systems, providing object-level inputs for subsequent tasks such as hazardous area intrusion recognition, worker–machinery distance assessment, unsafe behavior identification, and real-time safety warning.

5.2. Qualitative Analysis Under Challenging Construction Scenarios

5.2.1. Visual Comparison of Small-Object Detection Results

Figure 7 presents a comparison of visual detection results under different construction scenarios. Overall, the four groups of images are all collected from wide-field construction sites, where small target scales, sparse target distributions, complex background structures, and partial occlusions are commonly observed. For construction safety monitoring, distant workers, small safety-related targets, and personnel activities around machinery are often associated with high safety risks. However, these targets occupy only limited pixel regions in the images and are easily affected by ground textures, rebar structures, construction machinery, and building backgrounds. Therefore, Figure 7 effectively reflects the practical difficulties of safety-related target detection in open construction scenes.

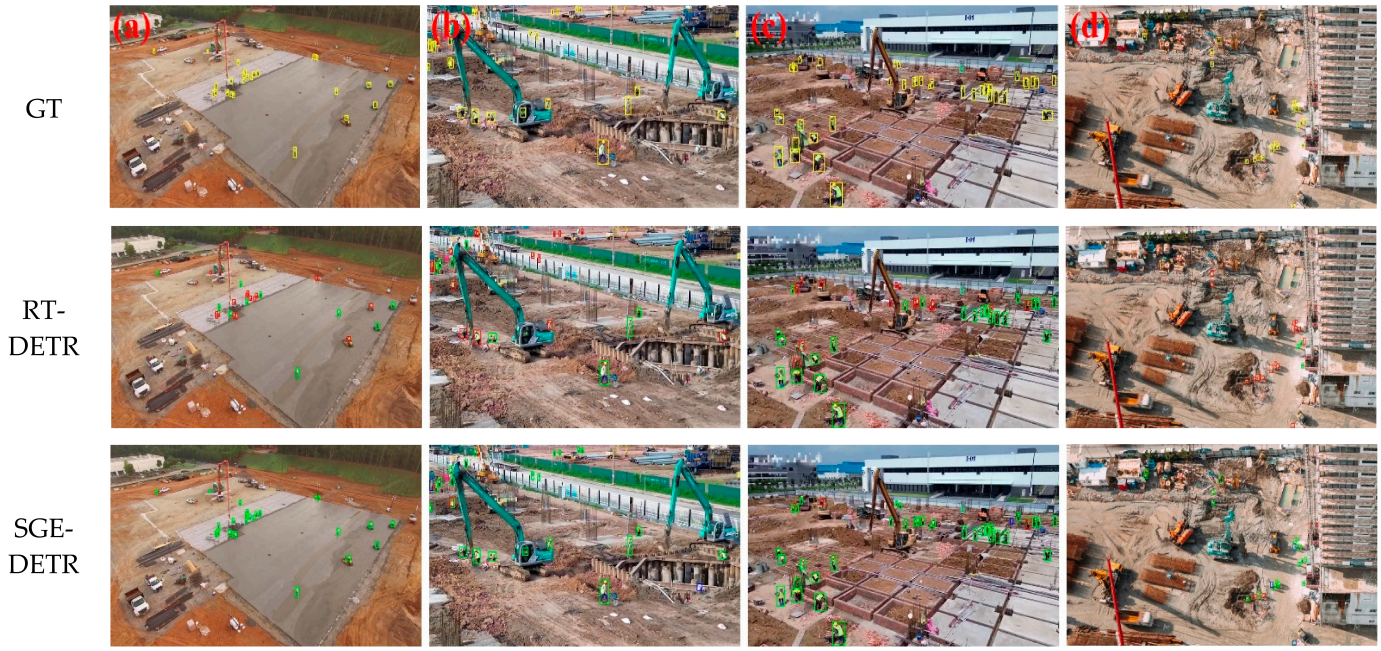


Figure 7. Qualitative comparison of small-object detection results in large open construction scenes. Columns (**a**–**d**) show different construction scenarios. Red, blue, and green boxes denote FN, FP, and TP, respectively.

From the overall results, RT-DETR can detect some salient targets, but it still suffers from missed detections and unstable localization for small-scale targets, distant targets, and targets in complex background regions. For example, in Figure 7a, the targets are small and sparsely distributed. RT-DETR shows weak responses to some distant workers or small safety-related targets, and the number of detected boxes is lower than that of the ground truth, indicating its limited capability in perceiving low-resolution small targets. In contrast, the proposed method preserves more small-target detection results, and the target

distribution is closer to the ground truth, demonstrating its stronger perception capability for weak and small targets in wide-area construction monitoring images.

In Figure 7b,c, construction machinery, fences, rebar structures, and complex ground textures appear simultaneously, resulting in obvious background interference. Such regions are usually of particular concern in safety monitoring, because potential safety risks are more likely to occur when workers are close to machinery, structural edges, or construction materials. Although RT-DETR can identify some workers and targets near equipment, it still produces several missed detections, and some detection boxes exhibit localization offsets or incomplete responses. By comparison, the proposed method performs more stably around machinery, structural junctions, and distant working areas, and can more sufficiently detect small-scale safety-related targets. This indicates its stronger modeling capability for weak-texture targets and edge details under complex background conditions.

Figure 7d shows a typical high-altitude top-view construction scene, which contains large machinery, earthwork areas, roads, buildings, and numerous small targets. The target scales differ significantly in this scene. RT-DETR is easily disturbed by complex background textures and large-scale structures, causing many small targets to be ignored and leading to obvious missed detections. In contrast, the proposed method produces more sufficient responses to small-scale targets, especially in dense working areas and distant small-target regions, where more valid detection boxes are obtained. This demonstrates its stronger feature representation and target discrimination capabilities in construction scenes where multi-scale targets coexist.

Overall, the proposed method shows better small-target detection capability and multi-scale adaptability in complex construction scenes. In particular, for distant workers, personnel around machinery, small safety-related targets, and targets located in regions with complex background textures, the proposed method effectively reduces missed detections and improves localization stability. These results indicate that the proposed method can better preserve fine-grained features of safety-critical small targets and enhance target discrimination under complex backgrounds, thereby providing more reliable visual detection support for risk perception in wide-field construction safety monitoring.

5.2.2. Heatmap Comparison of Attention and Gating Mechanisms

To further differentiate SGRC3 from existing attention and gating mechanisms and examine model responses under challenging construction-site conditions, we compared heatmap visualizations generated by different methods in three representative scenarios: high-angle monitoring, cluttered structural backgrounds with partial occlusion, and low-light scenes. As shown in Figure 8, Squeeze-and-Excitation (SE) [68] and Efficient Channel Attention (ECA) [69] highlight some target-related regions, but their activations remain relatively diffuse. Because these modules mainly recalibrate channel-wise feature responses, broad activations are still observed around building materials, rebars, formwork, and other textured background regions. This suggests that channel attention alone may be insufficient to separate weak safety-related targets from structured construction backgrounds. Gated convolution [70] introduces spatial modulation and reduces some irrelevant background responses; however, its activations for distant or weakly visible targets remain less stable, and some background structures are still highlighted.

By contrast, SGRC3 produces more compact activations around workers and safety-related local regions while suppressing irrelevant background textures more effectively. In the high-angle scene, the increased camera height reduces the apparent size of workers and weakens their local details; nevertheless, SGRC3 focuses more clearly on small-scale workers and reduces responses from surrounding material-stacking areas. In the cluttered scene, rebars, vertical components, and partially occluded targets introduce strong

background interference, whereas SGRC3 better suppresses repetitive structural responses while preserving target-related activations. In the low-light scene, reduced contrast and blurred object boundaries weaken the responses of all methods, but SGRC3 still maintains relatively clearer activations around visible safety-related targets.

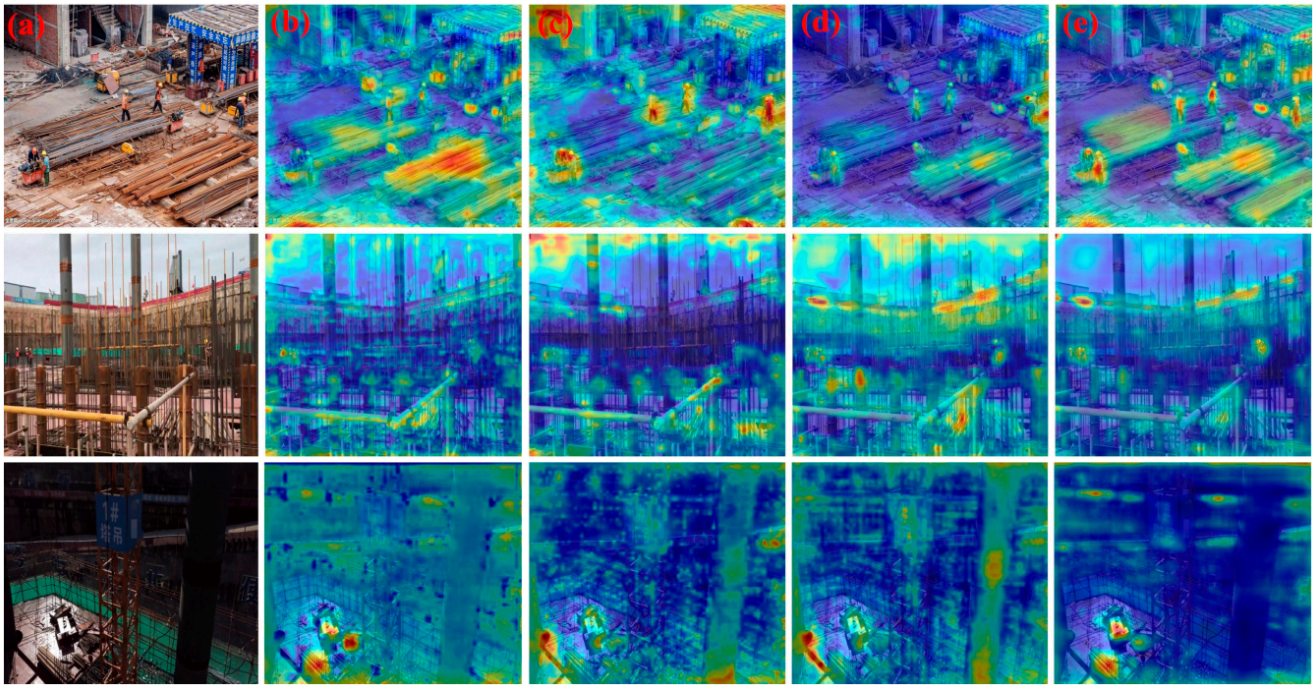


Figure 8. Heatmap comparison of different attention and gating mechanisms under challenging construction scenarios. (**a**) Images; (**b**) Baseline + SE; (**c**) Baseline + ECA; (**d**) Baseline + Gated convolution; (**e**) Baseline + SGRC3. The baseline model is RT-DETR-r18.

These observations also provide practical guidance for model deployment. High camera positions mainly increase the risk of missed detections by reducing target scale; cluttered backgrounds and occlusions increase the risk of false responses by introducing target-like edges; and low illumination weakens the reliability of edge and texture cues. Therefore, practical applications should consider appropriate camera placement, illumination enhancement, multi-frame verification, and scene-specific threshold adjustment to improve monitoring stability.

Overall, the heatmap results show that SGRC3 differs from SE and ECA by emphasizing spatially selective refinement rather than relying solely on channel-wise recalibration. Compared with generic gated convolution, SGRC3 embeds spatial gating into a lightweight reparameterized C3 structure, thereby improving local feature refinement and the suppression of background interference. However, extremely low illumination, severe occlusion, and very weak target boundaries may still weaken target responses and should be further investigated in future deployment-oriented studies.

5.3. Comparison with Related Work

Compared with recent RT-DETR-based improvements, SGE-DETR is motivated by a distinct design rationale. Many existing RT-DETR improvements primarily focus on enhancing general detection performance through backbone replacement, attention-mechanism optimization, feature-interaction reconstruction, or lightweight encoder–decoder modifications [71,72]. These strategies can improve general detection accuracy or inference efficiency, but they do not explicitly model the visual degradation patterns in large open construction scenes. In contrast, SGE-DETR is designed to address three challenges specific to large

open construction scenes: weak edge cues of distant safety-related targets, substantial scale differences between near-field and distant objects, and structured background interference from scaffolding, reinforcing bars, machinery, formwork, and material piles. Therefore, its advantages lie not only in architectural efficiency but also in scene adaptability.

More specifically, SREM enhances the weak contours and local structures of small safety-critical targets and differs from general edge-enhancement operations by introducing residual edge modulation within a scale-aware context. This design helps preserve weak target boundaries while reducing the risk of over-amplifying irrelevant background textures. DCAF adaptively balances multi-level features based on contextual and edge-density cues, rather than uniformly aggregating features as conventional FPN/PAN-style fusion structures do. SGRC3 suppresses redundant background textures while optimizing safety-related spatial responses, distinguishing it from channel-attention modules such as SE and ECA, as well as from generic gated convolutions. Instead of merely recalibrating channel weights or applying a general gating operation, SGRC3 embeds spatial gating into a reparameterized C3 structure to improve local feature selection under structured construction backgrounds. Together, these designs make SGE-DETR more suitable for wide-area construction safety perception than general RT-DETR improvement methods.

Compared with recent small-object detection methods, SGE-DETR also demonstrates more task-specific advantages. Existing studies on small-object detection generally improve detection performance by increasing feature resolution, strengthening multi-scale fusion, introducing attention mechanisms, or exploiting contextual information [41–45]. Small-object detectors such as STW-YOLO and SOD-YOLO further improve the detection performance of small or densely distributed targets through YOLO-based architectural modifications [54,60]. However, these methods mainly focus on enhancing small-object representations while paying less attention to the risk of amplifying construction background textures. In wide-area construction images, fine-grained structures are not always reliable target cues because background elements may produce similar edge and texture responses. SGE-DETR addresses this issue by combining edge-aware representation, density-guided adaptive fusion, and spatially gated refinement. Overall, SGE-DETR achieves a better balance among accuracy, robustness, and inference speed for safety monitoring in large open construction scenes.

5.4. Limitations and Future Work

Although the proposed SGE-DETR achieves favorable detection performance on the SODA and STWDs and shows certain advantages in small-object detection, multi-scale object adaptation, and target discrimination under complex backgrounds, this study still has several limitations.

1. The proposed method achieved satisfactory performance on a high-performance workstation, requiring only 14.50 M parameters and 32.50 GFLOPs, with both values lower than those of the RT-DETR-r18 baseline model and several Transformer-based alternatives. This indicates its preliminary potential for edge deployment. However, real construction sites usually lack high-performance GPU devices; therefore, models are typically required to run on resource-constrained devices. In addition, construction-site video streams are often affected by camera vibration, network transmission delay, motion blur, variations in installation height, and changes in monitoring viewpoints, which may weaken the model's detection stability and alarm reliability in continuous monitoring scenarios. Therefore, the model's embedded deployment reliability, long-term operational performance, and real-time response capability in real-world engineering scenarios still require further validation. Future work will consider deploying SGE-DETR to construction-site video monitoring systems, tower crane

cameras, or edge computing devices using deployment techniques such as TensorRT acceleration, FP16 inference, single-batch streaming inference, and model pruning or quantization. Its engineering applicability will then be tested using continuous video streams in terms of complex environments, long-term operation, latency, power consumption, thermal stability, and alarm reliability.

2. Although this study achieved good detection performance on typical wide-area construction-scene datasets such as SODA and STWD, the experiments were mainly conducted offline using public datasets. Adverse visual conditions on real construction sites may still affect model stability. Factors such as heavy dust, dense fog, rainy weather, and low nighttime illumination can reduce image contrast, blur target contours, and introduce noise or pseudo-edge structures. Because SREM enhances small-object representation by modulating edge, contour, and local structural responses, its edge-guided mechanism may be sensitive to the clarity and reliability of visual boundaries. Therefore, in these degraded scenarios, unreliable edge cues may cause SREM to overemphasize background contours or non-target structures, thereby increasing the risk of false alarms and reducing detection stability. Future work will further investigate adverse-weather data augmentation, noise-robust feature learning, and robust edge-gating strategies to improve the model's reliability on real construction sites.
3. This study focuses on improving the visual detection of safety-critical targets, such as workers and PPE-related objects, thereby providing a reliable object-level perception basis for construction safety warnings. However, practical construction safety monitoring is not limited to worker and PPE detection; it also involves other risk events, such as falling objects, unauthorized access, and equipment abnormalities or failures. The proposed edge-preserving and background-suppression strategies may be extendable to these tasks. For example, in falling-object detection, edge-preserving representations may help enhance the local contours of small tools, material fragments, or falling components, although multi-frame motion trajectories and velocity changes are still needed to confirm such events. For unauthorized-access detection, the detection results can be integrated with electronic fences, hazardous-area modeling, or BIM-based scene information to determine whether workers have entered restricted zones. For equipment abnormality or failure monitoring, background suppression and local structure enhancement may provide useful visual cues for identifying abnormal equipment postures, missing protective components, or visible appearance anomalies. Therefore, future research could build upon the proposed detection framework by further integrating multi-object tracking, hazardous-area modeling, worker–machinery spatial relationship analysis, multi-frame confirmation mechanisms, and multi-source sensing information, thereby extending its applicability to comprehensive construction safety warning systems.
4. Future work can also be extended to UAV inspection and wide-area monitoring scenarios with large fields of view. UAV inspection can flexibly cover temporary work areas and high-risk regions that are difficult for fixed cameras to observe continuously. However, targets in UAV images usually exhibit more significant scale variations, while viewpoint changes and background disturbances are also more complex. In future studies, SGE-DETR can be combined with UAV video inspection, construction-site electronic fencing, BIM scene information, and real-time alarm systems to further enhance its application value in dynamic inspection, hazardous area intrusion recognition, and intelligent safety warning on construction sites.

6. Conclusions

This study proposes SGE-DETR, a safety-oriented detection framework for wide-area construction safety perception. The framework is designed to address the difficulty of detecting safety-critical small targets in large open construction scenes, where distant workers, PPE-related objects, and personnel near machinery often appear with limited pixels, weak textures, and complex background interference. By integrating scale-aware residual edge modulation, density-guided adaptive feature fusion, and spatially gated reparameterized feature refinement, SGE-DETR enhances fine-grained target representation, improves multi-scale feature interaction, and reduces redundant background responses.

Experiments on the SODA and STWDs show that the proposed method improves detection performance across different object scales, particularly for dense and distant small workers, while maintaining relatively low model complexity and computational cost. The ablation results further indicate that these gains mainly stem from more effective feature representation and cross-scale fusion rather than from increased model size.

The findings suggest that wide-area construction safety perception requires more than general object detection capability. Reliable monitoring in large open scenes depends on preserving weak edge and structural cues, adapting to severe scale variations, and suppressing interference from construction backgrounds. In this regard, SGE-DETR provides a feasible visual perception approach for safety-critical small-object detection in scenarios such as high-view surveillance, tower-crane cameras, perimeter monitoring, and UAV-based inspection.

Future work will further evaluate SGE-DETR using continuous video streams and under real construction-site deployment conditions, and will integrate it with tracking, danger-zone modeling, personnel–machinery spatial relationship analysis, and multi-frame confirmation strategies for practical safety warning systems.

Author Contributions: L.S.: Methodology, Data curation, Conceptualization, Writing—Original draft preparation. Y.S.: Supervision, Funding acquisition, Writing—Reviewing and Editing, Project administration. H.L.: Software, Validation, Formal analysis, Writing—Reviewing and Editing. Z.G.: Data collection, Investigation, Resources. D.N.: Data curation, Visualization, Investigation. X.Y.: Data collection, Validation, Resources. K.G.: Software, Formal analysis, Data processing. Y.L.: Validation, Resources, Data collection. Y.W.: Conceptualization, Methodology, Writing—Reviewing and Editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This study was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement: The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Zhou, W.; Zhao, L.; Huang, H.; Wang, C. Data-Efficient Object Detection on Construction Sites Using Reweighting Mechanism and Cross-Batch Contrastive Learning. *IEEE Trans. Ind. Inf.* **2025**, *21*, 6028–6037. [CrossRef]
2. Chen, Z.; Li, J.; Li, Z.; Zhang, H.; Kong, L. ConstrID: A Hybrid Benchmark of Real and Synthetic Images for Construction Worker Re-Identification. *Ain Shams Eng. J.* **2026**, *17*, 104299. [CrossRef]
3. Seo, J.; Han, S.; Lee, S.; Kim, H. Computer Vision Techniques for Construction Safety and Health Monitoring. *Adv. Eng. Inform.* **2015**, *29*, 239–251. [CrossRef]
4. Chen, P.; Nie, X.; Ning, Y.; Zhang, Y. Learning Efficient and Adaptive Cross-Channel Dependencies for Weakly-Supervised Object Detection. *IEEE Trans. Multimed.* **2025**, *27*, 8954–8966. [CrossRef]
5. Weng, Y.; Yang, K.; Liu, Z.; He, W.; Tang, X. LGVLM-mIoT: A Lightweight Generative Visual-Language Model for Multilingual IoT Applications. *IEEE Internet Things J.* **2025**, *12*, 13311–13326. [CrossRef]

6. Zaidi, S.F.A.; Yang, J.; Abbas, M.S.; Hussain, R.; Lee, D.; Park, C. Vision-Based Construction Safety Monitoring Utilizing Temporal Analysis to Reduce False Alarms. *Buildings* **2024**, *14*, 1878. [CrossRef]
7. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Rose, T.M.; An, W. Detecting Non-Hardhat-Use by a Deep Learning Method from Far-Field Surveillance Videos. *Autom. Constr.* **2018**, *85*, 1–9. [CrossRef]
8. Zhong, W.; Rasouli, S.; Singh, A.K.; Mohandes, S.R.; Antwi-Afari, M.F.; Cheung, C.; Manu, P.; Agrawal, U. The Adoption of UAVs for Enhancing Safety in Construction Industry: A Systematic Literature Review. *Intell. Infrastruct. Constr.* **2025**, *1*, 3. [CrossRef]
9. Jacob-Loyola, N.; Muñoz-La Rivera, F.; Herrera, R.F.; Atencio, E. Unmanned Aerial Vehicles (UAVs) for Physical Progress Monitoring of Construction. *Sensors* **2021**, *21*, 4227. [CrossRef] [PubMed]
10. Song, H. Multi-Scale Safety Helmet Detection Based on RSSE-YOLOv3. *Sensors* **2022**, *22*, 6061. [CrossRef] [PubMed]
11. Wang, S.; Park, S.; Kim, J.; Kim, J. Safety Helmet Monitoring on Construction Sites Using YOLOv10 and Advanced Transformer Architectures with Surveillance and Body-Worn Cameras. *J. Constr. Eng. Manag.* **2025**, *151*, 04025186. [CrossRef]
12. Lee, J.; Lee, S. Construction Site Safety Management: A Computer Vision and Deep Learning Approach. *Sensors* **2023**, *23*, 944. [CrossRef] [PubMed]
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Proceedings of the Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Las Vegas, NV, USA, 2016; pp. 779–788. [CrossRef]
16. Long, X.; Cui, W.; Zheng, Z. Safety Helmet Wearing Detection Based on Deep Learning. In Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 15–17 March 2019; pp. 2495–2499. [CrossRef]
17. Nath, N.D.; Behzadan, A.H.; Paal, S.G. Deep Learning for Site Safety: Real-Time Detection of Personal Protective Equipment. *Autom. Constr.* **2020**, *112*, 103085. [CrossRef]
18. Li, X.; Hu, M.; Li, B.; Tong, R. OAM-YOLO: A Real-Time Small Object Detection Framework for PPE Compliance Monitoring in Industrial Environments. *Process Saf. Environ. Prot.* **2025**, *204*, 108058. [CrossRef]
19. An, X.H.; Zhou, L.; Liu, Z.G.; Wang, C.Z.; Li, P.F.; Li, Z.W. Dataset and Benchmark for Detecting Moving Objects in Construction Sites. *Autom. Constr.* **2021**, *122*, 103482. [CrossRef]
20. Roy, T.; Zhou, Y. From Agricultural Waste to Building Insulation: A Chemical-Free Composite from Banana Stem and Jackfruit Seed Starch. *Bioresour. Technol.* **2026**, *452*, 134591. [CrossRef] [PubMed]
21. 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]
22. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229. [CrossRef]
23. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2021**, arXiv:2010.04159. [CrossRef]
24. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605. [CrossRef]
25. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. [CrossRef]
26. Ding, S.; Li, J.; Wang, M. Optimizing the Influence of Fiber Orientation on the Mechanical Properties of Ultrahigh-Performance Concrete via Data Augmentation and Machine Learning. *Struct. Concr.* **2026**. [CrossRef]
27. Li, Z.; Liu, J.; Xu, Y.; Lee, L.-H.; Zhao, Z. A Computational Study on the Hydrodynamics of Real-Time Pitch Manipulation in Dolphins during Self-Propulsion: The Promoting Effect of Caudal Fin under Offset Sinusoidal Waveform. *Phys. Fluids* **2025**, *37*, 111905. [CrossRef]
28. Chen, D.; Xu, C.; Yu, J.; Wang, Q.; Fang, H.; Yin, S.; Lookman, T.; Chen, R. Interpretable Machine Learning Framework for Nb–Si Based Alloy Design with Enhanced Fracture Toughness. *Adv. Sci.* **2026**, e75815. [CrossRef] [PubMed]
29. Yan, J.; Zhou, N.; Cheng, Y.; Zhang, F.; Wang, H.; Wang, M.; Jin, B.; Li, M.; Lu, Q.; Zhang, W. Application of Machine-Vision-Driven Physics-Informed Neural Networks in Pantograph–Catenary System State Detection. *Mech. Syst. Signal Process.* **2026**, *257*, 114577. [CrossRef]

30. Wang, Q.-A.; Huang, X.-Y.; Ni, Y.-Q.; Feng, W.-S.; Wang, J.-F.; Qiu, S.; Zhou, P.; Li, J.-P.; Luo, Y. Uncertainty-Aware, High-Precision Multi-Step Prediction of Structural Health Monitoring Sensor Streams under Extreme Typhoon Events: An Enhanced Bayesian Dynamic Linear Model Leveraging the Kernel Regression Basis Function for Severe Environmental Adaptation. *Measurement* **2026**, *262*, 120050. [CrossRef]
31. Li, Z.; Yan, Y.; Zhao, Z.; Xu, Y.; Hu, Y. Numerical Study on Hydrodynamic Effects of Intermittent or Sinusoidal Coordination of Pectoral Fins to Achieve Spontaneous Nose-up Pitching Behavior in Dolphins. *Ocean Eng.* **2025**, *337*, 121854. [CrossRef]
32. Chen, Z.-H.; Xu, Z.-D.; Entezami, A.; Sarmadi, H.; Gao, W.; Behkamal, B. Dynamic Monitoring of Bridges Subjected to Freezing-Induced Variability by Kernel-Enhanced Deep Anomaly Detection. *Eng. Struct.* **2026**, *362*, 122971. [CrossRef]
33. Li, J.; Chen, Z.; Li, Z.; Kong, L.; Zhang, H. Integrating Computer Vision and Audio Signals for Concrete Vibration Activity Recognition and Assessment. *J. Constr. Eng. Manag.* **2026**, *152*, 04026099. [CrossRef]
34. Han, Y.; Li, Z.; Li, J. Pavement Condition Detection Using Acceleration Data Collected by Smartphones Based on 1D Convolutional Neural Network. *JCE* **2024**, *76*, 979–991. [CrossRef]
35. Wu, J.; Cai, N.; Chen, W.; Wang, H.; Wang, G. Automatic Detection of Hardhats Worn by Construction Personnel: A Deep Learning Approach and Benchmark Dataset. *Autom. Constr.* **2019**, *106*, 102894. [CrossRef]
36. Han, K.; Zeng, X. Deep Learning-Based Workers Safety Helmet Wearing Detection on Construction Sites Using Multi-Scale Features. *IEEE Access* **2022**, *10*, 718–729. [CrossRef]
37. Hayat, A.; Morgado-Dias, F. Deep Learning-Based Automatic Safety Helmet Detection System for Construction Safety. *Appl. Sci.* **2022**, *12*, 8268. [CrossRef]
38. Jiao, X.; Li, C.; Zhang, X.; Fan, J.; Cai, Z.; Zhou, Z.; Wang, Y. Detection Method for Safety Helmet Wearing on Construction Sites Based on UAV Images and YOLOv8. *Buildings* **2025**, *15*, 354. [CrossRef]
39. Fang, W.; Ding, L.; Zhong, B.; Love, P.E.D.; Luo, H. Automated Detection of Workers and Heavy Equipment on Construction Sites: A Convolutional Neural Network Approach. *Adv. Eng. Inform.* **2018**, *37*, 139–149. [CrossRef]
40. Wang, S. Domain-Adaptive Faster R-CNN for Non-PPE Identification on Construction Sites from Body-Worn and General Images. *Sci. Rep.* **2026**, *16*, 4793. [CrossRef] [PubMed]
41. Li, J.; Zhao, X.; Zhou, G.; Zhang, M. Standardized Use Inspection of Workers' Personal Protective Equipment Based on Deep Learning. *Saf. Sci.* **2022**, *150*, 105689. [CrossRef]
42. Park, M.; Tran, D.Q.; Bak, J.; Park, S. Small and Overlapping Worker Detection at Construction Sites. *Autom. Constr.* **2023**, *151*, 104856. [CrossRef]
43. Song, C.; Li, Y. A YOLOv8 Algorithm for Safety Helmet Wearing Detection in Complex Environment. *Sci. Rep.* **2025**, *15*, 24236. [CrossRef] [PubMed]
44. Lai, C.-H.; Wu, T.-E.; Wang, C.-C. Enhancing Information Security in Smart Manufacturing Through Least Significant Bit Steganography in Engineering Drawings. *J. Comput. Inf. Sci. Eng.* **2025**, *25*, 091006. [CrossRef]
45. Li, J.; Xie, S.; Zhou, X.; Zhang, L.; Li, X. Real-Time Detection of Coal Mine Safety Helmet Based on Improved YOLOv8. *J. Real-Time Image Process.* **2025**, *22*, 26. [CrossRef]
46. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef] [PubMed]
47. Feng, Q.; Xu, X.; Wang, Z.; Feng, Q.; Xu, X.; Wang, Z. Deep Learning-Based Small Object Detection: A Survey. *MBE* **2023**, *20*, 6551–6590. [CrossRef] [PubMed]
48. Xiuling, Z.; Huijuan, W.; Yu, S.; Gang, C.; Suhua, Z.; Quanbo, Y. Starting from the Structure: A Review of Small Object Detection Based on Deep Learning. *Image Vis. Comput.* **2024**, *146*, 105054. [CrossRef]
49. Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. VisDrone-DET2021: The Vision Meets Drone Object Detection Challenge Results. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; IEEE: Montreal, BC, Canada, 2021; pp. 2847–2854. [CrossRef]
50. Hua, W.; Chen, Q. A Survey of Small Object Detection Based on Deep Learning in Aerial Images. *Artif. Intell. Rev.* **2025**, *58*, 162. [CrossRef]
51. Zhang, H.; Li, J.; Li, Z.; Kong, L.; Zhao, X. Real-Time Safety Helmet Detection Using Multi-Scale Feature-Gated Fusion. *Autom. Constr.* **2026**, *186*, 106891. [CrossRef]
52. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2017**, arXiv:1612.03144. [CrossRef]
53. Wei, J.; Chen, Z.; Zhang, Y.; Zhang, W.; Li, J.; Silva Lora, E.E.; Venturini, O.J.; Kozlov, A.N.; Isa, Y.M. An Adjustable Heat-to-Power Ratio Combined Heat and Power System Integrating Biomass Gasification, Solid Oxide Fuel Cell, and Gas Turbine: Energy, Exergy, Economic, and Exergoeconomic (4E) Analyses. *Int. J. Hydrogen Energy* **2026**, *244*, 155706. [CrossRef]
54. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768. [CrossRef]

55. Li, Z.; Xia, D.; Lei, M.; Yan, H. Numerical Investigation on Self-Propelled Hydrodynamics of Squid-like Multiple Tentacles with Synergistic Expansion. *Ocean Eng.* **2023**, *287*, 115808. [CrossRef]
56. Liu, J.; Luo, H.; Liu, H. Deep Learning-Based Data Analytics for Safety in Construction. *Autom. Constr.* **2022**, *140*, 104302. [CrossRef]
57. Li, Y.; Zhang, X.; He, Z.; Xie, J.; Zhang, Y.; Wang, J.; Deng, J. Response of Sludge EPS under the Action of Coupled Alternating Electric Field-Ammonium Persulfate Dewatering Technology: Structural Changes and Mechanism Analysis. *Chem. Eng. Sci.* **2026**, *323*, 123267. [CrossRef]
58. Liu, H.-I.; Tseng, Y.-W.; Chang, K.-C.; Wang, P.-J.; Shuai, H.-H.; Cheng, W.-H. A DeNoising FPN with Transformer R-CNN for Tiny Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4704415. [CrossRef]
59. 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]
60. Yin, Y.; Liu, X. Small Target Worker Detection Based on Improved YOLOv12 for Large Construction Scenes. *Autom. Constr.* **2026**, *182*, 106722. [CrossRef]
61. Sohan, M.; Sai Ram, T.; Rami Reddy, C.V. A Review on YOLOv8 and Its Advancements. In *Proceedings of the Data Intelligence and Cognitive Informatics*; Jacob, I.J., Piramuthu, S., Falkowski-Gilski, P., Eds.; Springer Nature: Singapore, 2024; pp. 529–545. [CrossRef]
62. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011. [CrossRef]
63. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
64. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524. [CrossRef]
65. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788. [CrossRef]
66. Li, Y.; Li, Q.; Pan, J.; Zhou, Y.; Zhu, H.; Wei, H.; Liu, C. SOD-YOLO: Small-Object-Detection Algorithm Based on Improved YOLOv8 for UAV Images. *Remote Sens.* **2024**, *16*, 3057. [CrossRef]
67. Zheng, P.; Zhao, Y.; Cui, Z.; Li, Y. PRNet: Original Information Is All You Have. *arXiv* **2025**, arXiv:2510.09531. [CrossRef]
68. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [CrossRef]
69. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Seattle, WA, USA, 2020; pp. 11531–11539. [CrossRef]
70. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T. Free-Form Image Inpainting with Gated Convolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 4470–4479. [CrossRef]
71. Li, S.; Zhou, Y.; Wang, M.; Liu, L.; Wu, T. MRC-DETR: A High-Precision Detection Model for Electrical Equipment Protection in Power Operations. *Sensors* **2025**, *25*, 4152. [CrossRef] [PubMed]
72. Liu, Z.; Sun, C.; Wang, X. DST-DETR: Image Dehazing RT-DETR for Safety Helmet Detection in Foggy Weather. *Sensors* **2024**, *24*, 4628. [CrossRef] [PubMed]

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.