symmetry

MDPI

Article

YOLO-GP: A Multi-Scale Dangerous Behavior Detection Model Based on YOLOv8

Bushi Liu [1], Cuiying Yu [1,*], Bolun Chen [1,2,†] and Yue Zhao [1,†]

[1] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, China; bsliu@hyit.edu.cn (B.L.); chenbolun@hyit.edu.cn (B.C.); zhaoyue1876@hyit.edu.cn (Y.Z.)
[2] Department of Physics, University of Fribourg, CH-1700 Fribourg, Switzerland
* Correspondence: yucuiying@hyit.edu.cn
† These authors contributed equally to this work.

Abstract: In recent years, frequent chemical production safety incidents in China have been primarily attributed to dangerous behaviors by workers. Current monitoring methods predominantly rely on manual supervision, which is not only inefficient but also prone to errors in complex environments and with varying target scales, leading to missed or incorrect detections. To address this issue, we propose a deep learning-based object detection model, YOLO-GP. First, we utilize a grouped pointwise convolutional (GPConv) module of symmetric structure to facilitate information exchange and feature fusion in the channel dimension, thereby extracting more accurate feature representations. Building upon the YOLOv8n model, we integrate the symmetric structure convolutional GPConv module and design the dual-branch aggregation module (DAM) and Efficient Spatial Pyramid Pooling (ESPP) module to enhance the richness of gradient flow information and the capture of multi-scale features, respectively. Finally, we develop a channel feature enhancement network (CFE-Net) to strengthen inter-channel interactions, improving the model's performance in complex scenarios. Experimental results demonstrate that YOLO-GP achieves a 1.56% and 11.46% improvement in the mAP@.5:.95 metric on a custom dangerous behavior dataset and a public Construction Site Safety Image Dataset, respectively, compared to the baseline model. This highlights its superiority in dangerous behavior object detection tasks. Furthermore, the enhancement in model performance provides an effective solution for improving accuracy and robustness, promising significant practical applications.

Keywords: dangerous behavior detection; multi-scale features; gradient information enhancement; complex scene detection

Citation: Liu, B.; Yu, C.; Chen, B.; Zhao, Y. YOLO-GP: A Multi-Scale Dangerous Behavior Detection Model Based on YOLOv8. *Symmetry* **2024**, *16*, 730. https://doi.org/10.3390/sym16060730

Academic Editors: Boris Malomed and Alice Miller

Received: 15 May 2024
Revised: 6 June 2024
Accepted: 6 June 2024
Published: 12 June 2024

1. Introduction

In recent years, with the rapid development of the economy, chemical enterprises have become an integral part of modern society. However, along with the rapid growth of chemical enterprises, significant safety hazards have been brought to people's social lives [1]. These enterprises produce a wide range of chemicals used in daily life. However, the production process of these chemicals often involves high temperatures, high pressures, and flammable or explosive materials, which could lead to serious casualties and environmental pollution in the event of accidents. Ensuring the safety of chemical enterprises and preventing accidents is crucial. In chemical production, various dangerous behaviors such as failure to wear personal protective equipment, violation of regulations, and non-compliance with prescribed procedures are often among the main causes of accidents. Therefore, the accurate monitoring and identification of dangerous behaviors in chemical enterprises have become an urgent need. By accurately identifying and monitoring dangerous behaviors of chemical workers during the production process, timely warnings can be issued, and measures can be taken to effectively reduce the probability of accidents, ensuring the safety of personnel and the integrity of production facilities.

The monitoring of dangerous behaviors is of great significance in the safety domain. For example, behaviors such as operating without wearing helmets, smoking, or using phones during operations by workers may pose serious safety risks. These behaviors not only endanger personal safety but may also affect the safety and stability of the surrounding personnel and work environments.

However, traditional methods of monitoring dangerous behaviors often rely on manual inspection, which is inefficient, costly, and subjective. Given this issue, technologies in computer vision and deep learning offer new opportunities to address this challenge. In recent years, significant progress has been made in target detection technology, particularly with the application of deep learning models. Both two-stage target detection algorithms like Region-based Convolutional Neural Network (R-CNN), Fast Region-based Convolutional Neural Network (Fast-RCNN), and Faster Region-based Convolutional Neural Network (Faster-RCNN) [2–4], and one-stage target detection algorithms such as You Only Look Once (YOLO) and Single-Shot Multi-Box Detector (SSD) [5,6] have been widely applied in dangerous behavior detection. Many studies have optimized both two-stage and one-stage object detection models. In two-stage approaches, Mask R-CNN, by introducing segmentation branches into the detection framework, not only detects objects but also performs instance segmentation, significantly improving detection accuracy [7]. Cascade R-CNN introduces multi-stage regression and classification modules to progressively refine detection results, particularly excelling in handling high-quality detection frameworks [8]. Although two-stage object detection methods have advantages in accuracy, they exhibit lower detection efficiency compared to one-stage methods. Therefore, one-stage object detection methods are more suitable for domains requiring timely responses, such as hazardous behavior detection. However, the accuracy of one-stage methods is slightly lower. Consequently, many current studies are focused on optimizing one-stage object detection methods to enhance their accuracy. For instance, by introducing multi-scale feature fusion networks [9,10] and attention mechanisms [11–14] into the YOLO model, detection accuracy and speed can be significantly improved. Additionally, EfficientDet [15], through optimized network architectures and scaling strategies, achieves higher accuracy while maintaining high efficiency. These advancements not only enhance the detection capabilities of the models but also strengthen their performance in handling complex backgrounds and multi-scale targets. Nevertheless, one-stage detection models still tend to suffer from false positives and misses, especially when facing scenes with high complexity and multiple objects of different scales. These issues are particularly prominent in models not specifically designed for hazardous behavior detection tasks.

To address these challenges, this study adopts the current state-of-the-art, highly accurate single-stage object detection algorithm YOLOv8n as the baseline model for further improvements. YOLOv8n, a specific variant of YOLOv8, is characterized by its lightweight nature, making it suitable for resource-constrained applications. This research focuses on detecting whether workers wear safety helmets, use mobile phones, and smoke. The choice of YOLOv8n aims to enhance detection accuracy and provide a more reliable solution for real-world safety surveillance applications. Consequently, this study proposes the YOLO-GP algorithm, an improved version of YOLOv8n, and successfully applies it to dangerous behavior detection. The network architecture proposed in this study makes the following four contributions:

- In this study, an innovative symmetric structure of grouped pointwise convolutional (GPConv) is designed, which enhances the model's feature representation and expressiveness by integrating feature fusion in the channel dimension and combining various feature extraction methods.
- A dual-branch aggregation module (DAM) is designed to replace the C2f module of the original model to obtain richer gradient flow information, to solve the problem of the baseline model's poor accuracy in locating dangerous behavioral targets and its poor ability to discriminate between small targets such as smoking and phone usage.

- By fusing the innovative efficient spatial pyramid pooling (ESPP) module to the neck of the model, we can effectively improve the recognition ability of dangerous behaviors through multi-scale feature capture and feature fusion so that the model can accurately understand and differentiate dangerous behavior targets involving different scales.
- To solve the problem of insufficient channel correlation, which leads to the poor performance of the model in detecting dangerous behaviors in complex scenes, a channel feature enhancement network (CFE-Net) is designed to enable the model to better understand the interactions between different channels, to achieve the purpose of improving the accuracy of the model in detecting dangerous behaviors in complex scenes.

2. Related Work

In recent years, the field of dangerous behavior detection has been devoted to exploring various technological approaches to comprehensively enhance workplace safety. These approaches encompass a range of methods, including traditional algorithms and deep learning algorithms. The core focus of research typically revolves around utilizing surveillance data to identify workers' dangerous behaviors, such as the absence of safety helmets, smoking, and mobile phone usage during work in chemical enterprises. However, existing studies tend to lean towards specific types of dangerous behaviors, posing challenges in addressing the diverse requirements of actual workplace scenarios.

To gain a more comprehensive understanding of the application of dangerous behavior detection technology in different work settings, this study focuses on summarizing the latest research findings in the field of safety helmet wearing, smoking, and mobile phone usage detection both domestically and internationally. The aim is to elucidate the overall development trends in this field and provide valuable insights into the comprehensive application of dangerous behavior detection technology in practical workplace settings through an in-depth exploration of various technological approaches. This endeavor not only holds the potential to improve the overall safety of workplaces but also aids in effectively reducing the potential risks faced by chemical enterprises.

2.1. Traditional Methods for Dangerous Behavior Detection Algorithm

In the field of dangerous behavior detection, traditional methods primarily focus on utilizing computer vision and image processing techniques, as well as conventional feature engineering methods. For instance, Abu H. et al. proposed a detection method for automatically detecting helmets to ensure construction safety. They first combined the frequency domain information of images with Histograms of Oriented Gradients (HOGs) for detecting construction workers and then applied a combination of color-based and Circular Hough Transform (CHT) feature extraction techniques to detect the usage of helmets by construction workers [16]. Seshadri et al. introduced a computer vision-based driver mobile phone usage detection system. By employing facial landmark tracking algorithms, the system can automatically identify whether the driver brings the phone close to the ear. The research validated the system using challenging Strategic Highway Research Program (SHRP2) facial view videos, demonstrating its effectiveness under natural driving conditions. By combining direct methods and various features, the system achieved satisfactory performance on facial pose verification data, providing new insights into understanding driver behavior [17]. Wang et al. proposed a method for phone behavior detection using semi-supervised Support Vector Machine (SVM) models. Although the method involves a large iteration amount during the detection process, leading to slow data processing speed and real-time issues, it offers a unique approach to determining phone usage behavior [18]. Pan et al. proposed a method combining Gaussian Mixture Models (GMMs) and frame differencing for extracting features of regions of interest and analyzing smoking behavior through RGB color features. However, this method is influenced by factors such as movement speed and weather in smoking behavior detection and suffers from low real-time performance and accuracy issues [19]. Ai Bo utilized a combination of Gaussian Mixture Models with background subtraction to extract foreground object

information, then performed HOG feature extraction on regions of interest, and determined the presence of smoking behavior in the current frame through a classifier. However, this traditional method is prone to be affected by factors such as weather and pedestrian speed in smoking behavior detection, limiting its accuracy and real-time performance [20].

In summary, traditional methods are often constrained by manually designed features and rules, which may limit their performance in addressing complex scenes and diverse dangerous behaviors. However, with the rise in deep learning, methods based on deep learning have gradually achieved significant breakthroughs in the field of dangerous behavior detection.

2.2. Dangerous Behavior Detection Based on Deep Learning

Currently, both domestically and internationally, there is active exploration of the application of deep learning technology in safety behavior detection in the chemical industry, especially in detecting dangerous behaviors during chemical workers' operations. Deep learning technology has been widely proven to have tremendous potential in reducing workplace accident risks and improving work efficiency. In the field of dangerous behavior detection, object detection techniques based on deep learning play a crucial role. Object detection methods are mainly divided into single-stage and two-stage approaches, each with its unique characteristics and advantages when it comes to performing dangerous behavior detection tasks [21].

First, the two-stage target detection method accomplishes the target detection task through two stages. It first generates a series of candidate regions through a region proposal network and then performs target detection on these regions. The two-stage target detection algorithm usually has an advantage in accuracy and is suitable for dangerous behavior detection scenarios that require high detection accuracy. In the field of dangerous behavior detection, many domestic and foreign researchers are committed to using two-stage target detection algorithms to achieve the purpose of detecting whether there are potential dangerous behaviors in the work of workers through the in-depth analysis of picture data.

For example, Dey et al. proposed a context-driven detection method for distracted driving using in-vehicle cameras. This method employs a novel computer vision technique to detect distracted driving by identifying and analyzing objects like hands and smartphones inside the vehicle. By its unique context-driven approach, it provides real-time feedback regarding the specific reasons for distraction, thereby enhancing driving safety [22]. Senyurek et al. introduced a deep learning algorithm utilizing a convolutional neural network (CNN) and long short-term memory network (LSTM) architecture to detect smoking behavior from respiratory signals. Compared to traditional feature-based classification frameworks, the advantage of the CNN-LSTM model lies in learning appropriate features from respiratory inductive plethysmography (RIP) sensor signals through the CNN layer, providing superior performance for smoking detection [23]. Han et al. proposed a method for fast smoking behavior detection. Firstly, the face area is taken as the scope of smoking detection, effectively reducing the detection area of smoking targets by utilizing the characteristic that human body targets are relatively large compared to the face area. Then, the Faster R-CNN model is used to determine whether smoking behavior exists [24]. Wang et al. proposed a method for identifying unsafe behaviors of construction workers based on text mining and image recognition technology, divided into three stages. Firstly, a deep learning algorithm is used to identify the safety equipment of construction workers. Secondly, the classification and detection of unsafe behaviors are completed through Faster R-CNN. Finally, in the third stage, the identification and tracking of personnel in dangerous areas are conducted, achieving comprehensive recognition of unsafe behaviors of construction workers on construction sites [25]. Chen et al. presented a real-time automatic detection system for safety helmet-wearing based on the Faster R-CNN algorithm. The improved algorithm introduces Retinex image enhancement technology, effectively overcoming interference from factors such as light and distance. This technology

improves the quality of images in complex outdoor scenes of substations, enabling timely and effective detection of individuals not wearing safety helmets, and providing reliable safety monitoring for substation construction [26].

Second, single-stage target detection methods directly predict the location and category of targets in the input image without explicitly generating candidate regions. Typical single-stage target detection algorithms include YOLO, SSD, etc. These algorithms are characterized by high real-time performance, simplicity, and high efficiency, which are especially suitable for scenarios with high real-time requirements in dangerous behavior detection. In the field of dangerous behavior detection, many scholars at home and abroad also widely apply single-stage target detection methods to achieve timely and efficient detection of dangerous behavior.

For instance, Aboah et al. proposed a real-time multi-class helmet violation detection method using few-shot data sampling techniques and YOLOv8. They extracted frames with partially different backgrounds from a large number of video frames and applied data augmentation operations to these extracted frames using test-time augmentation strategies. Finally, they trained and tested the YOLOv8 model, achieving real-time detection goals [27]. Fan et al. introduced a helmet-wearing detection method based on the EfficientDet algorithm. They first optimized the initial clustering centers using the K-Means++ clustering algorithm, then introduced the SeparableConv2D network. They combined the Simple and Efficient Bi-directional Feature Pyramid Network (BiFPN) proposed in the EfficientDet algorithm to extract image feature maps. They utilized the Channel Correlation Loss (CC-Loss) function as the classification loss function to constrain specific relationships between classes and channels, maintaining separability within and between classes, thereby improving the accuracy of the model detection [28]. Yang et al. proposed a deep learning-based SSD algorithm for detecting illegal driving behaviors. The detection of driver-driving behaviors mainly includes using mobile phones, smoking, and not wearing seat belts. Utilizing the SSD algorithm can effectively address the issue of whether the driver is violating driving regulations during the driving process, significantly reducing the occurrence of traffic accidents [29]. Zhao et al. presented a smoking behavior detection method for drivers based on the Feature Pyramid Network (FPN). By combining the FPN and dilated convolution technology, they detected small objects in driver images and identified their smoking behavior [30]. She et al. proposed an improved YOLOx-based algorithm for small target smoking detection. By adding an attention mechanism module to focus on global information in the feature extraction network and concentrating attention within the target area through scale addition, they increased the use of deep networks. They also optimized the loss function by replacing it with the Generalized Intersection over Union (GIoU) loss function, addressing the shortcomings of IoU [31].

Although deep learning has made some progress in dangerous behavior detection, challenges remain, such as insufficient detection effectiveness and relatively low accuracy. These issues also exist in the job safety and security scenes addressed in this study, especially in situations involving multiple target occlusions and environmental interference, rendering existing methods impractical. Compared to existing research, this study pays full attention to the characteristics of job safety and security scenes, addressing the issue of low detection accuracy of multi-scale targets in complex scenes in dangerous behavior target detection tasks.

3. Material and Methods

3.1. YOLO-GP Algorithm Overview

In 2023, Ultralytics released the YOLOv8 algorithm, which significantly improved detection accuracy and real-time performance while reducing network parameters compared to previous versions [32]. When facing practical scenarios such as chemical plant production workshops, this study chose the lightweight version of YOLOv8n as the base network and made improvements to it. It is noteworthy that YOLOv8 retains the Cross-Stage Partial Network (CSP) concept, Path Aggregation Network (PANet), and Spatial Pyramid Pooling

Fast (SPPF) module from YOLOv5 while integrating many excellent techniques from the real-time object detection field.

The network architecture depicted in Figure 1 is the enhanced network based on the baseline model: YOLO-GP. The YOLO-GP detector comprises three main components: Backbone, Neck, and Head. The Backbone serves as the foundational component of the model, employing a lightweight structure to enhance efficiency when handling large-scale data. Additionally, it incorporates the DAM to enrich gradient flow information, addressing the challenges of target localization and small target identification faced by the original model. Furthermore, it employs the innovative ESPP module to improve the recognition capability of dangerous behaviors. The Neck component further refines the model by enhancing channel correlation modeling through the CFE-Net network, effectively improving dangerous behavior detection performance in complex scenarios. Finally, the Head section is responsible for generating detection results, including object categories and positional information.

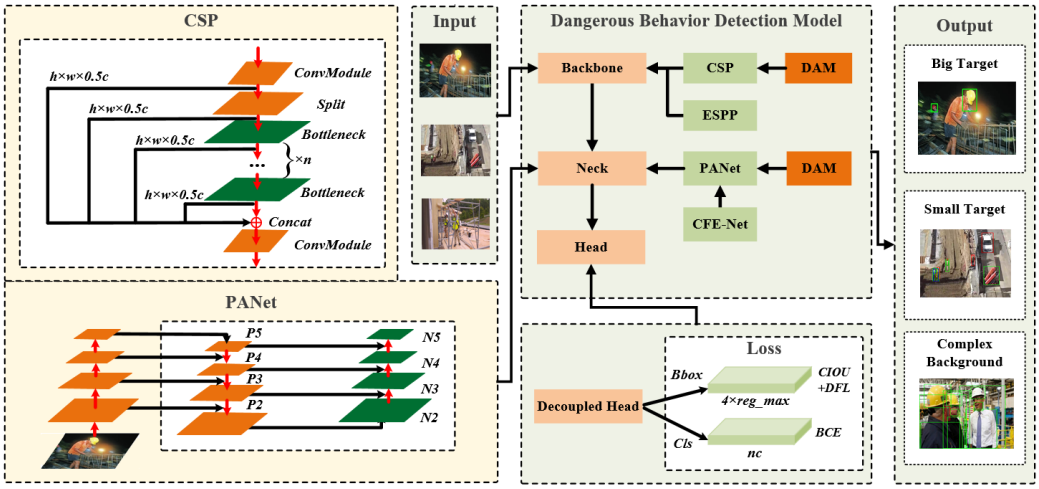

Figure 1. YOLO-GP algorithmic framework. The original image is input from the Input part, while it is detected by the Dangerous Behavior Detection Model, and finally the targets under Big Target, Small Target and Complex Background are output from the Output part.

3.2. Improvement Strategies

3.2.1. Grouped Pointwise Convolutional

The GPConv module, primarily composed of group convolution [33] and pointwise convolution [34], is a convolutional module with a symmetric structure. Specifically, group convolution is a variant of the convolution operation that divides input channels into several groups, with each group's channels convolving only with channels from the same group. This partitioning of convolution kernel parameters into multiple groups allows each group to convolve only with a subset of input channels. More precisely, if the number of input channels is C and they are divided into G groups, then each group contains C/G channels. This method of combining results from different groups according to certain rules effectively reduces the model's parameters and computational complexity while also enhancing the network's representational capacity. By grouping input channels, channels within each group only convolve with other channels in the same group, facilitating interaction between different channels and aiding in the extraction of richer and more diverse features.

Pointwise convolution is a convolution operation with a kernel size of 1×1, where it considers only the value of a single pixel at each position of the input. It is primarily used for linear combinations across the channel dimension. Although it does not have a spatial receptive field, it enhances the network's representational capacity and performance by performing linear transformations and feature fusion along the channel dimen-

sion. A comparison between group convolution and pointwise convolution is illustrated in Figure 2.

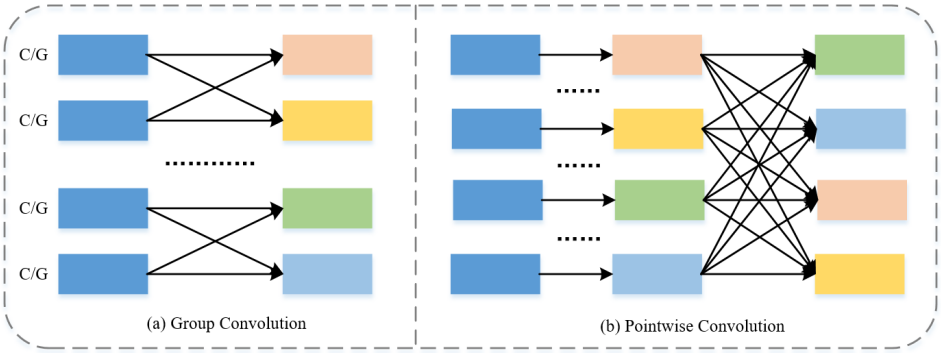

Figure 2. Comparison between group convolution and pointwise convolution. (**a**) Schematic diagram of group convolution: shows the basic principle and the way of action of group convolution operation. (**b**) Schematic diagram of pointwise convolution: shows the basic principle and the way of action of the pointwise convolution operation.

By leveraging the combination of group convolution and pointwise convolution, the GPConv module can fully exploit different feature extraction approaches, thus enhancing the diversity and richness of features. Moreover, by integrating the advantages of group convolution and pointwise convolution, the GPConv module can better utilize features at different levels. Through feature fusion in the channel dimension and combining different feature extraction methods, the GPConv module possesses richer feature representation and higher expressive power.

Furthermore, Leaky-ReLU is introduced as the activation function, combined with batch normalization for normalization, to accelerate training convergence and improve model stability. The structure diagram of the GPConv module is illustrated in Figure 3.

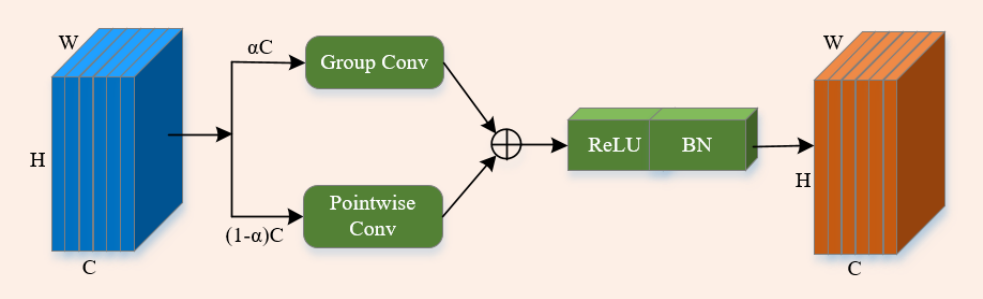

Figure 3. Diagram illustrating the structure of the GPConv module. ReLU denotes Leaky-ReLU activation function, BN denotes batch normalization.

The GPConv module processes the input features through two branches separately. Specifically, for the input $X \in R^{N \times C_{in} \times H_{in} \times W_{in}}$, N is the batch size, C_{in} is the number of input channels, and H_{in} and W_{in} are the height and width of the input. After the group convolution and pointwise convolution operations, the convolution operation between the input tensor X and the group convolution kernel tensor W_{gc} and the pointwise convolution kernel tensor W_{pwc} is performed:

$$Z_{gc} = \sum_{i=1}^{g} \left(X_i * W_{gc_i} \right), \tag{1}$$

$$Z_{pwc} = X * W_{pwc}, \tag{2}$$

where W_{gc_i} represents the weight tensor of the group convolution kernel, X_i represents the i-th group of the input tensor, and W_{pwc} represents the weight tensor of the pointwise convolution kernel. Then, the two tensors Z_{gc} and Z_{pwc} are elementwise added to obtain

the sum feature map tensor Z_{sum}. Subsequently, the resulting tensor is passed through an activation function and normalization to obtain the final output Y. The specific formulas are as follows:

$$Z_{sum} = Z_{gc} + Z_{pwc}, \tag{3}$$

$$A_{LR} = LeakyRelu(Z_{sum}, \alpha), \tag{4}$$

$$Y(A_{LR}) = \gamma * \frac{A_{LR} - \mu}{\sqrt{\delta^2 + \varepsilon}} + \beta, \tag{5}$$

where, α represents the negative slope of the activation function, γ and β are learnable parameters, μ and δ^2 represent the mean and variance of the feature map Z, and ε prevents division by zero error.

3.2.2. Dual-Branch Aggregation Module

The YOLOv8 algorithm model retains the Cross-Stage Partial concept introduced in YOLOv5 [35]. It splits the feature map of the forward propagation into two parts, divides them along the channel dimension, and introduces cross-stage connections between the two divided parts, enabling the interaction and partial fusion of forward and backward feature maps. This structure maintains detection accuracy while accelerating convergence speed and reducing computational complexity.

Although the CSP structure introduces cross-stage connections to facilitate interaction between forward and backward feature maps, the transmission of connected information may be limited. Especially when spanning multiple stages, the gradient propagation may be hindered, resulting in insufficient gradient flow information, which affects the accurate localization of targets, especially smaller targets (such as smoking and phone usage) in the dangerous behavior dataset. At the same time, the design of the CSP structure pursues simplicity and effectiveness but may sacrifice certain information processing capabilities.

To solve the above problems, we propose a DAM module. This module combines the advantages of feature segmentation and cross-stage connection in the CSP structure and realizes more comprehensive feature extraction and information interaction by adopting the method of dual-branch aggregation. At the same time, the structural features of GPConv are utilized to provide the module with richer feature characterization capability and higher expressive power. The structure of DAM is shown in Figure 4.

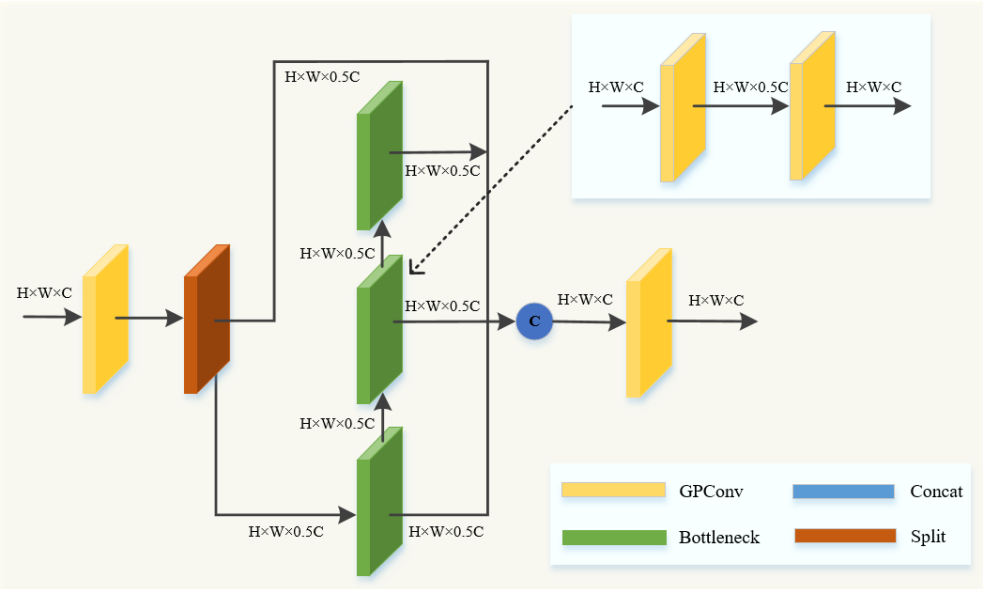

Figure 4. Schematic diagram of DAM structure.

The DAM consists of GPConv and multiple Bottleneck modules, with each Bottleneck module being composed of two GPConv modules in series. The overall structure of DAM involves processing the input through a GPConv module once, and then splitting the output features into two parts. One part undergoes a direct cross-stage connection, while the other part is processed through multiple repeated Bottleneck modules. This stacking approach enables the network to become deeper and extract richer feature representations, thus making the network more effective in handling complex tasks.

Finally, the output features from both parts are concatenated and further processed through a GPConv module, thereby enhancing the model's representational power. This design fully leverages the advantages of the GPConv module and Bottleneck module to better extract feature information, allowing the model to more accurately locate dangerous behavior targets and improve its ability to locate small targets such as smoking or phone usage. The design of the DAM module enables the model to obtain richer gradient flow information, effectively addressing the issue of the poor localization accuracy of dangerous behavior targets in the baseline model.

3.2.3. Efficient Spatial Pyramid Pooling

The traditional Spatial Pyramid Pooling (SPP) structure is a pooling method designed to address the issue of varying input sizes. In conventional convolutional neural networks, fully connected layers typically require inputs of fixed sizes, which is inconvenient for handling images of different sizes. The SPP structure introduces multi-scale pooling operations, allowing the network to accept images of any size and generate fixed-length feature vectors [36]. The design of SPPF aims to improve the efficiency and speed of spatial pyramid pooling, thereby accelerating the inference process of object detection models. This enhanced spatial pyramid pooling technique helps improve the detection performance of models on objects of different scales while maintaining accuracy and speeding up inference.

Although the SPPF structure has achieved good results in speeding up inference, it may have limitations when dealing with images containing rich spatial information. Particularly when handling objects with multiscale and complex structures, the SPPF structure may not fully utilize both local and global information of features, resulting in less rich and accurate feature representations.

To overcome this limitation, we propose an ESPP module to further optimize the feature extraction process. Compared to the traditional SPP structure, the ESPP module combines spatial pyramid pooling with group pointwise convolution modules. It conducts local feature extraction on the feature maps at different scales and performs group pointwise convolution operations in the channel dimension, thereby obtaining more representative feature representations. The structure diagram of the ESPP module is shown in Figure 5.

In the YOLO-GP network model, the features extracted from the Backbone part are fused through the ESPP module to integrate features from different levels. Firstly, the input data undergo feature extraction and transformation through 1 × 1 and 3 × 3 convolutional layers, followed by a GPConv module. This module internally includes group convolution, pointwise convolution, and batch normalization, effectively enhancing the feature representation capability. Subsequently, the extracted features are fed into multiple sizes (5 × 5, 9 × 9, and 13 × 13) of max-pooling layers to capture diversified feature information at different scales. Then, a series of convolution operations further process and fuse these features to enhance their representation capability. Simultaneously, another input branch undergoes feature processing through the GPConv module. Finally, the features extracted from these two branches are merged and comprehensively processed through the final convolutional layer to generate the ultimate output for object detection. This network structure has the capability of multi-scale feature capture and feature fusion. By employing multi-scale max-pooling layers, feature fusion, and multi-path feature information processing, this module helps extract multi-scale semantic information, enhancing the model's perception of objects at different scales. Particularly in cases where the size of objects varies greatly or local information is highly important for determining object significance, the ESPP module

can provide richer feature representations, thus enhancing the detection performance of multi-scale objects.

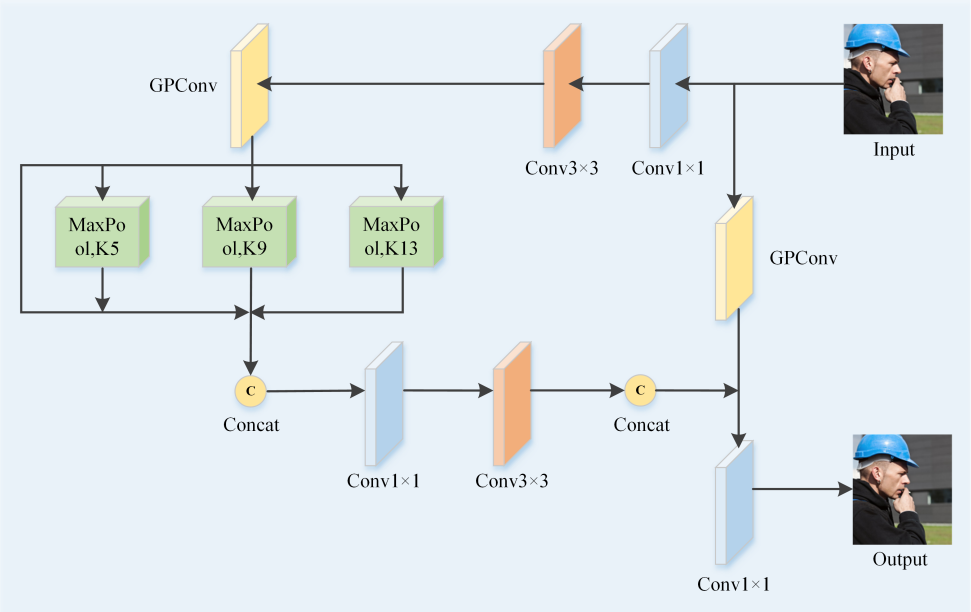

Figure 5. ESPP module structure diagram.

In summary, the ESPP module combines the advantages of spatial pyramid pooling and group pointwise convolution, enabling better adaptation to features of different scales and complexities, thereby improving the model's detection effectiveness of target objects. Its capabilities in multi-scale feature capture and feature fusion are expected to yield good results in practical applications.

3.2.4. Channel Feature Enhancement Network

In the task of detecting dangerous behaviors, a single frame image may contain multiple different dangerous behaviors simultaneously. This situation can lead to occlusion, resulting in missed detections and false alarms. To address this issue, our study integrates the CFE-Net into the neck of the baseline model to enhance the model's focus on the importance of channels and thus improve its performance.

The CFE-Net is an innovative convolutional neural network structure designed to enhance object detection performance and improve the model's perception of key features. The core of the CFE-Net lies in the fusion of GPConv with squeeze-and-excitation operations. The GPConv module emphasizes the modeling of branch–channel relationships through dual-branch convolution operations, allowing the network to capture image structure information more finely. Additionally, to further enhance the model's focus on crucial information, we introduce squeeze-and-excitation operations to adaptively adjust the weights of each channel [37]. Through global average pooling and a series of fully connected layers, dynamic attention weights are generated for each channel, enabling the network to focus more on the most informative channels in the image. Figure 6 shows the schematic diagram of CFE-Net.

Specifically, for a given input tensor $X \in R^{H \times W \times C}$, the convolution results of group convolution and pointwise convolution are separately processed by activation functions to obtain the activation tensors ${A_{gc}}' = \sigma({Z_{gc}}')$ and ${A_{wpc}}' = \sigma({Z_{wpc}}')$ for this part. Then, the two tensors are summed and normalized to obtain the output ${Y_c}'$ for this part. Subsequently, a squeeze operation is performed to average pool the feature map, resulting in a $1 * 1 * C$ vector, denoted as:

$$Z_c = \Gamma_{sq}(Y_c{}') = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Y_c{}'(i,j) \tag{6}$$

where $Z_{gc}{}'$ and $Z_{wpc}{}'$ represent the tensor representations of the group convolution branch and pointwise convolution branch, respectively, $A_{gc}{}'$ and $A_{wpc}{}'$ denote the tensors of the group convolution branch and pointwise convolution branch after activation function processing. $Y_c{}'$ indicates the output after batch normalization processing.

Following that, the vector Z_c obtained in the previous step is subjected to the excitation operation through fully connected layers F_1 and F_2, resulting in the target channel weight values W. After passing through two fully connected layers, different values in W represent the weight information for different channels, assigning different weights to each channel.

$$W = \Gamma_{ex}(z, F) = \mu(F_2 \delta(F_1 z)) \tag{7}$$

wherein $\delta(\cdot)$ represents the relu activation function, and $\mu(\cdot)$ represents the sigmoid activation function. After passing through two fully connected layers, a $1 * 1 * C$ vector is obtained. Then, through the Scalar Multiplication operation, which involves performing a multiplication operation between the feature map W and the input feature map $Y_c{}'$ corresponding to the channel, the output feature map Y is obtained:

$$Y = \Gamma_{sc}(Y_c{}', W) = Y_c{}' * W \tag{8}$$

In summary, CFE-Net emphasizes channel correlations through dual-branch convolutional operations and adaptively adjusts the weights of each channel using squeeze-and-excitation operations to focus the network on the most informative channels in the image. This enables the network to better understand the interactions between different channels, addressing the issue of insufficient channel correlations and thereby improving the model's performance in detecting dangerous behaviors in complex scenarios.

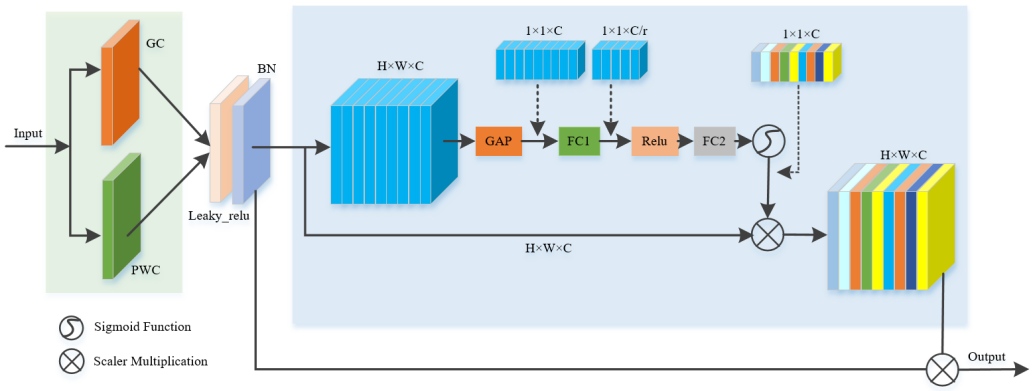

Figure 6. CFE-Net module structure diagram.

3.3. Experimental Environment and Parameter Settings

In this study, we utilized a high-performance computing platform based on the Windows 10 operating system. The hardware configuration includes an NVIDIA Tesla V100 SXM2 GPU with 16 GB of memory. To ensure the reliability of experimental data, all experiments were conducted under consistent hardware settings. PyTorch 1.8.0 was employed as the primary development framework, with CUDA 10.2.89 used for training acceleration. Python version 3.8 was utilized during training. We employed the stochastic gradient descent (SGD) optimizer for optimizing model parameters, with specific hyperparameters listed in Table 1.

During the experimental process, we found that compared to the 300 training epochs recommended by Ultralytics, optimal results could be achieved with 100 training epochs. Therefore, we set the training epochs for all models to 100. All experiments were conducted

in the same simulated environment to ensure the reliability of experimental data and the reproducibility of results. During training, the training time for each epoch varied depending on the size of the dataset and the complexity of the model. Smaller datasets and simpler models required shorter training times, while larger datasets and complex models required longer training times.

Table 1. Hyperparameter settings.

Hyperparameter	Values
Batch size	32
Epoch	100
NMS IoU	0.7
Initial Learning Rate	0.01
Final Learning Rate	0.01
Weight decay	0.0005
Momentum	0.937
Input size	640

3.4. Datasets

For the task of detecting dangerous behaviors among personnel in chemical plants, this study relies on a self-built dangerous behavior dataset (DBD). This dataset was created by searching for relevant images related to safety helmets, smoking, and phone usage through keyword searches and manual screening on the internet. When selecting images, we rigorously screened scenes and contexts involving dangerous behaviors such as not wearing a helmet, smoking, and using a phone to ensure the diversity and representativeness of the dataset. We then meticulously cleaned and filtered the combined collection using advanced image and video processing techniques, including deep learning, to eliminate noise and outliers. During the data cleaning process, we focused on the quality and clarity of the images to ensure the accuracy and reliability of the annotations. The final dataset contains a total of 5063 images, all annotated in YOLO format using Labelimg. Our detection tasks specifically focused on four dangerous behaviors: wearing a helmet ("helmet"), bare head ("head"), using a phone ("phone"), and smoking ("smoke"). During the annotation process, we strictly adhered to accurate and consistent annotation standards for marking dangerous behaviors in the images. To enhance the diversity and richness of the dataset, we employed various data augmentation techniques, including Mosaic augmentation, color augmentation (hue, saturation, and brightness), image flipping, translation, and scaling. These data augmentation techniques help improve the model's generalization ability and robustness. Finally, we divided the dataset into training, testing, and validation sets in a 6:2:2 ratio, with 3040 images for training and 2023 images for testing and validation. This dataset serves as a comprehensive resource for research experiments.

To validate the effectiveness of the proposed model, experiments were conducted on the Construction Site Safety Image Dataset (CSSID) [38]. This dataset consists of 2801 images covering various safety-related scenarios on construction sites. The labels include "safety helmet", "mask", "no safety helmet", "no mask", "no safety vest", "person", "safety cone", "safety vest", "machinery", and "vehicle", providing rich information for monitoring construction site safety. Due to the complexity of construction site environments, the dataset contains annotations of targets of different scales and sizes, enabling researchers to train and evaluate safety-related tasks at different scales and sizes. This dataset supports safety detection and monitoring tasks comprehensively. Partial image samples from both datasets are shown in Figure 7, while Figure 8 illustrates the distribution of categories in both datasets.

Figure 7. Selected images of the dataset are shown. (**a**) DBD partial images display. (**b**) CSSID partial images display.

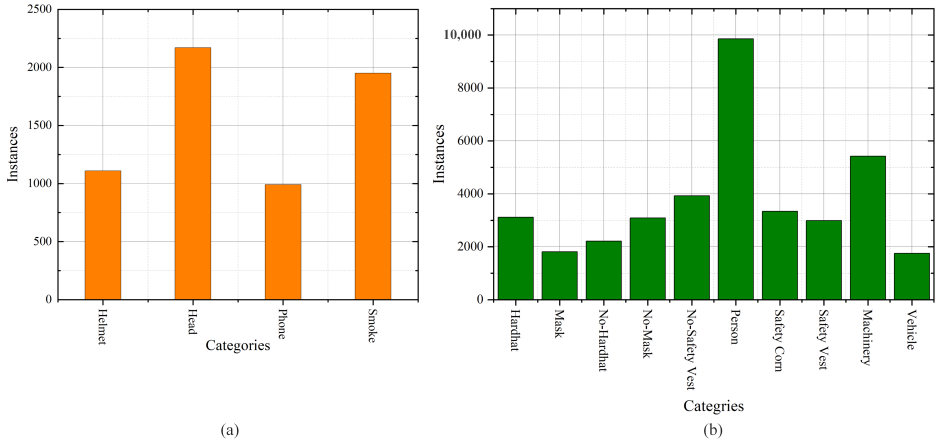

Figure 8. Schematic distribution of categories for the DBD and the CSSID. (**a**) Display of the number of target instances for the DBD. (**b**) Display of the number of target instances for the CSSID.

3.5. Evaluation Indicators

In this study, we employ Precision (P), Recall (R), F1-Score [39], and Mean Average Precision (mAP) [40] as metrics to evaluate the performance of our model. These metrics allow for a comprehensive assessment of the model's detection effectiveness and accuracy.

4. Experimental Results and Analysis

4.1. Ablation Experiment

To validate the effectiveness of each module on the model, this study conducted a series of ablation experiments on the Dangerous Behavior dataset and the Construction Site Safety Imagery dataset, respectively, and analyzed the results exhaustively. These ablation experiments were designed to systematically assess the impact of each component of the model on the overall performance. Consistency of the experimental parameters was maintained during the experiments to ensure comparable and reliable results. The performance evaluation metrics of each ablation experiment and the comparative results are detailed in Table 2.

In Table 2, we utilize the YOLOv8n model as the baseline model and integrate three innovative modules to enhance its performance, evaluating seven metrics. Additionally, we can observe more intuitively by combining the model curve comparison chart (as shown in Figure 9). Firstly, the DAM module resulted in a 1.68% and 0.9% improvement in mAP@.5 and mAP@.5:.95 on the DBD dataset, respectively, with a 1.3% increase in F1-Score. Meanwhile, there was a significant improvement of 5.86% and 11.46% in mAP@.5

and mAP@.5:.95 on the CSSID dataset. This indicates that DAM effectively enhances the detection performance by obtaining richer gradient flow information. The ESPP module led to an increase of 1.56% and 9.55% in mAP@.5:.95 on both datasets, demonstrating the effectiveness of this module. CFE-Net made additional improvements in the Neck part of the model, maintaining similar computational costs and parameter quantities on the DBD dataset while achieving a 1.16% increase in mAP@.5 and a 0.5% increase in F1-Score. Overall, the integration of these three modules constitutes the complete YOLO-GP model. Despite the increased computational costs of this model, there were improvements of 2.06%, 1.56%, 1.30%, 7.98%, 11.46%, and 3.8% in mAP@.5, mAP@.5:.95, and F1-Score on both datasets, respectively. These enhancements validate the effectiveness of the improvements and demonstrate the improvement in the accuracy and robustness of the model.

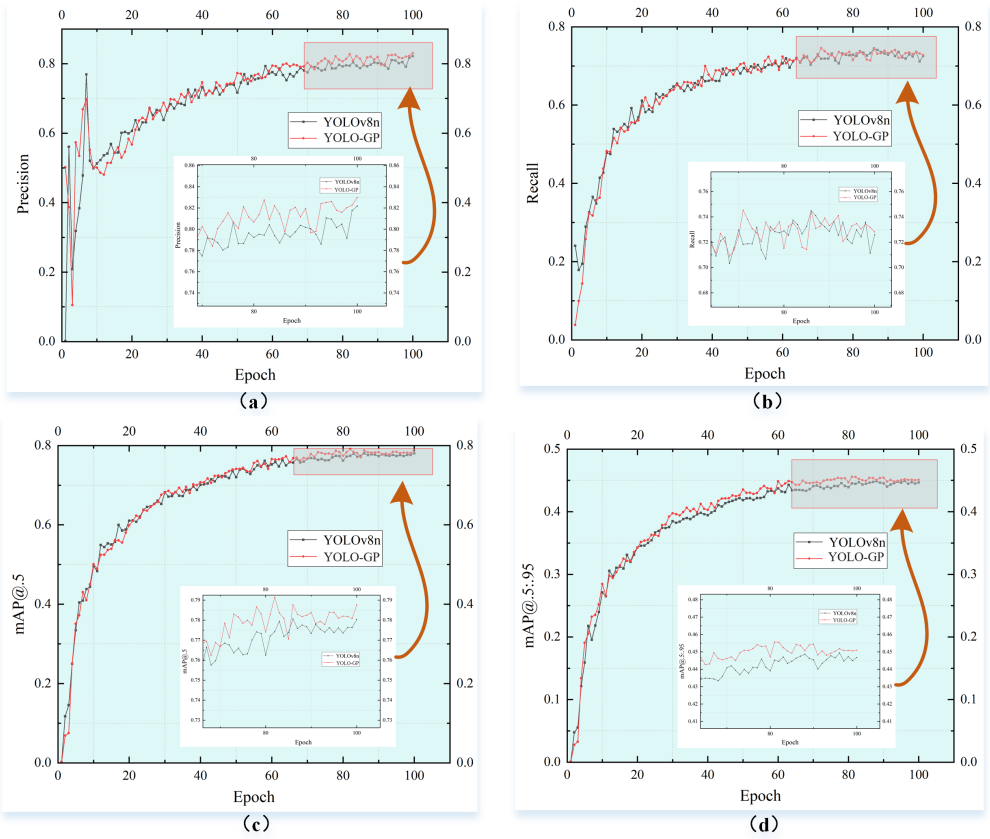

Figure 9. Comparison curves and detailed graphs of experiments conducted on the DBD dataset between the baseline model and the YOLO-GP model. (**a**) Precision comparison plot of the baseline model versus the YOLO-GP model; (**b**) Recall comparison plot of the baseline model versus the YOLO-GP model; (**c**) mAP@.5 comparison plot of the baseline model versus the YOLO-GP model; (**d**) mAP@.5:.95 comparison plot of the baseline model versus the YOLO-GP model.

Table 2. Ablation experiment.

Dataset	Models	Precision	Recall	mAP@.5	mAP@.5:.95	GFLOPs	F1-Score	Para/M
DBD	Baseline	80.1	73.3	77.6	44.9	8.9	76.5	3.16
	Baseline + DAM	84.1	72.4	78.9	45.3	13.7	77.8	5.49
	Baseline + ESPP	79.4	74.6	78.4	45.6	10.1	76.9	4.56
	Baseline + CFE-Net	80.3	74.0	78.5	45.4	15.1	77.8	3.38
	YOLO-GP	82.6	73.6	79.2	45.6	15.1	77.8	7.27

Table 2. *Cont.*

Dataset	Models	Precision	Recall	mAP@.5	mAP@.5:.95	GFLOPs	F1-Score	Para/M
CSSID	Baseline	80.6	54.6	61.4	31.4	8.9	65.1	3.16
	Baseline + DAM	85.4	57.0	65.0	35.0	13.7	68.4	5.49
	Baseline + ESPP	81.1	57.8	64.5	34.4	10.1	67.5	4.56
	Baseline + CFE-Net	72.8	53.7	58.6	30.4	10.1	61.8	3.38
	YOLO-GP	81.3	59.8	66.3	35.0	15.1	68.9	7.27

In summary, experiments with these two datasets revealed that the YOLO-GP algorithm has the following key advantages:

1. In the DBD dataset, the design of the DAM effectively enhances the Backbone and Neck structures of the YOLOv8n network, aiding in better gradient flow information capture. This helps improve the model's accuracy in localizing dangerous behavior targets, particularly addressing the challenge of discriminating small targets such as smoking or phone usage. In the CSSID dataset, this improvement also effectively addresses localization issues in complex construction site environments with diverse target scales;
2. On both datasets, the ESPP module enhances the model's ability to capture and fuse multi-scale features. This contributes to better recognition of dangerous behaviors, enabling the model to more accurately understand and differentiate targets involving various scales.
3. Integrating the CFE-Net enhances the model's understanding of inter-channel interactions, improving dangerous behavior detection performance in complex scenarios. In the DBD dataset, this helps address insufficient channel correlations, thereby better understanding the correlation between different dangerous behaviors. The CSSID dataset assists the model in handling various safety equipment, personnel, and objects present in construction site environments, enhancing the model's robustness and generalization capability.

By applying these three improvements to practical scenarios in two different datasets, the model demonstrates enhanced robustness and efficiency in diverse environments, improving its practicality and adaptability in safety detection and monitoring tasks.

4.2. Convergence Curve

In practical applications, ensuring that the model achieves high accuracy and fast convergence is crucial for its robustness and stability. Therefore, this study conducted 100 rounds of training and testing on both the baseline model and the YOLO-GP model, comparing their convergence curves. Figure 10 illustrates the convergence curves of classification loss (cls_loss) and localization loss (box_loss) for both the baseline model and the YOLO-GP model. The vertical axis represents the loss values during the network training process, while the horizontal axis represents the number of iterations of the network.

The experimental results indicate that both models initially exhibit relatively high loss values during the early stages of training. However, as training progresses, the loss values of both models gradually decrease and tend to converge, especially within the first 20 epochs. As training continues, the loss values of the networks continue to decrease, indicating that the networks gradually fit the training data. Overall, the YOLO-GP model demonstrates superior convergence performance and stability by maintaining lower loss values while ensuring convergence speed.

In summary, the experimental results validate the effectiveness of the YOLO-GP model improvements. The model achieves significantly improved detection accuracy, particularly for multi-scale objects. Furthermore, it exhibits better convergence performance and stability, making it suitable for various practical applications.

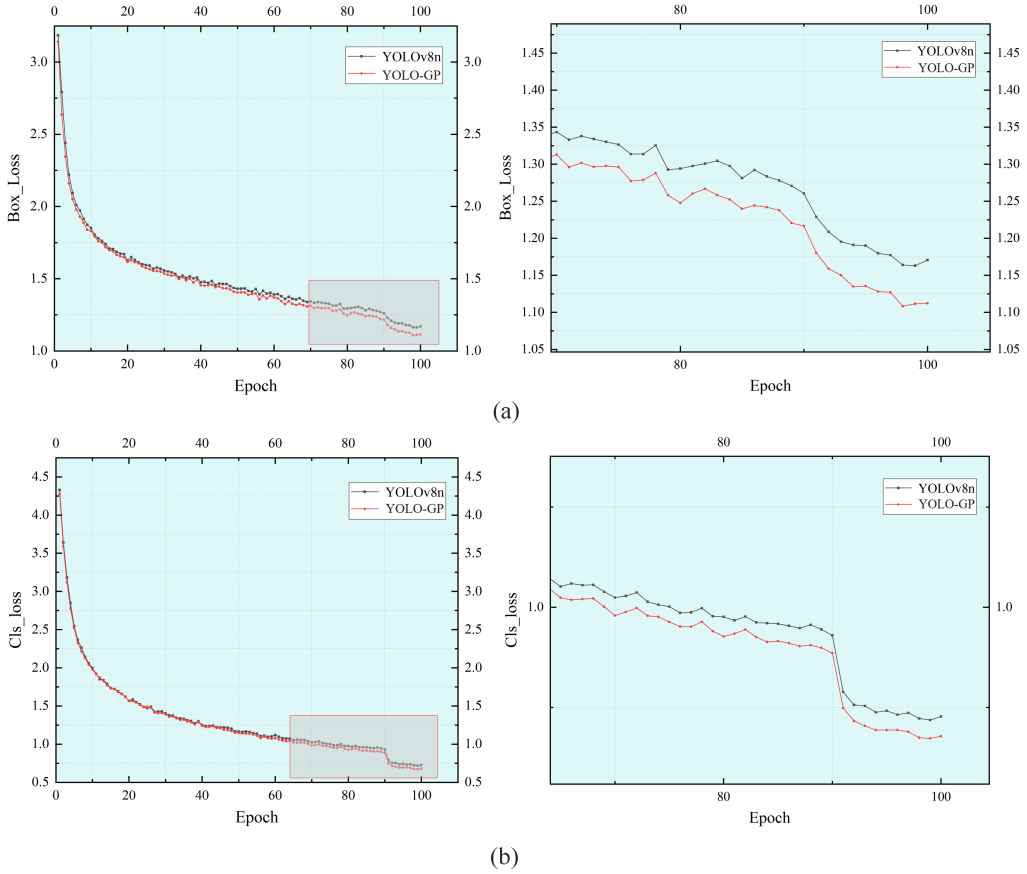

Figure 10. Comparison curves of classification loss and localization loss between the baseline model and the YOLO-GP model and their end-detail plots. (**a**) Cls_loss comparison curves. (**b**) Box_Loss Comparison Curves.

4.3. Activation Function Parameter Selection Experiment

In the Leaky ReLU activation function, the negative_slope parameter defines the slope when the input is negative. Typically, the ReLU activation function outputs zero for negative inputs, while Leaky ReLU allows negative inputs to pass through by multiplying them with a small slope instead of completely zeroing them out. Therefore, the negative slope parameter controls the magnitude of this slope.

For object detection models, the appropriate choice of the negative_slope parameter can influence the model's learning capability and convergence speed. A suitable negative_slope value can enable the model to better learn complex features and patterns, thus improving its stability and generalization ability. However, excessively large or small negative slope values may lead to unstable training or performance degradation. Therefore, selecting the appropriate negative_slope parameter is crucial for the performance of object detection models.

As shown in Table 3, through extensive experimentation, we found that setting the negative_slope parameter to 0.4 resulted in the best performance of the object detection model. Additionally, analyzing the basic principles of Leaky ReLU reveals its primary function of addressing the gradient vanishing problem associated with standard ReLU. When the negative slope is too small (e.g., 0.2 or 0.3), the gradient for negative inputs becomes very small. Although this mitigates part of the gradient vanishing issue, it can still result in excessively weak gradients, thereby affecting the efficiency of model training. Conversely, when the negative slope is too large (e.g., 0.5 or 0.6), the gradient for negative inputs becomes too large, which may lead to gradient explosion and unstable training. With a negative slope of 0.4, the gradient for negative inputs is neither too small nor too large, maintaining a balance in gradient propagation. This balance allows the model to train

stably and efficiently. Therefore, incorporating an appropriate negative slope in the Leaky ReLU activation function can significantly enhance the detection accuracy and stability of the model.

Table 3. Experiments on the effect of Negative_slope parameter size on CFE-Net.

Negative_Slope	Precision	Recall	mAP@.5	mAP@.5:.95
0.2	78.3	73.0	77.0	44.5
0.3	77.5	72.5	76.3	44.7
0.4	80.3	74.0	78.5	45.4
0.5	81.0	71.5	77.2	44.6
0.6	78.5	71.8	77.5	44.5

4.4. Validation of the Validity of the ESPP Module

To demonstrate the advantages of ESPP design in scale-aware object detection, this study conducted comparative experiments between ESPP and a series of popular SPP modules. The included SPP modules can directly replace the original SPPF method and include Simplified SPPF (SimSPPF), Spatial Pyramid Pooling Cross-Stage Partial Channel (SPPCSPC), Atrous Spatial Pyramid Pooling (ASPP), Receptive Field Block (RFB) [41–44], Spatial Pyramid Pooling Fast Cross-Stage Partial Channel (SPPFCSPC), and ESPP.

All hyperparameters and configurations in this study remained unchanged, and each module was replaced with SPPF on the baseline YOLOv8n model. Experiments were conducted on both the DBD and CSSID datasets, comparing six metrics: Precision, Recall, mAP@.5:.95, mAP@.5, parameter count, and GFLOPs. The performance impact of different SPP modules on the baseline model on the DBD and CSSID datasets is shown in Table 4.

Table 4. Comparative experimental results of DBD and CSSID datasets using multiple mainstream SPP modules.

Dataset	Method	Precision	Recall	mAP@.5	mAP@.5:.95	GFLOPs	Para/M
DBD	Baseline + SPPF	80.1	73.3	77.6	44.9	8.9	3.16
	Baseline + SPPCSPC	81.2	73.7	78.1	45.6	10.1	4.72
	Baseline + SPPFCSPC	78.2	72.8	77.5	45.2	10.1	4.77
	Baseline + SimSPPF	80.9	74.2	77.7	45.1	8.9	3.16
	Baseline + ASPP	77.6	75.8	77.9	45.2	10.5	5.22
	Baseline + RFB	80.8	72.6	77.3	44.7	9.0	3.32
	Baseline + ESPP	79.4	74.6	78.4	45.6	10.1	4.56
CSSID	Baseline + SPPF	80.6	54.6	61.4	31.4	8.9	3.16
	Baseline + SPPCSPC	78.8	58.6	64.2	34.6	10.1	4.72
	Baseline + SPPFCSPC	80.7	56.5	62.5	32.8	10.1	4.77
	Baseline + SimSPPF	72.4	56.2	60.7	30.7	8.9	3.16
	Baseline + ASPP	77.2	54.5	61.0	30.8	10.5	5.22
	Baseline + RFB	78.8	58.3	63.0	32.6	9.0	3.32
	Baseline + ESPP	81.1	57.8	64.5	34.4	10.1	4.56

From Table 4, it can be observed that ESPP outperforms other SPP modules significantly on both datasets. For the DBD dataset, compared to SPPF, the ESPP module improved the baseline model's mAP@.5 by 1.03 percentage points and mAP@.5:.95 by 1.56 percentage points. Compared to introducing SPPCSPC, SPPFCSPC, and ASPP into the baseline model, ESPP not only achieves higher detection accuracy but also has a lower parameter count and computational cost. Similar results are also evident in the training on the CSSID dataset. Compared to SPPF, DPSPP improved the baseline model's mAP@.5 and mAP@.5:.95 by 9.55 and 5.05 percentage points, respectively. Overall, DPSPP proves to be a more efficient module compared to various mainstream SPP modules.

The improvement in accuracy can be attributed to two main factors. Firstly, ESPP enhances the representation capability of feature maps through the structural characteristics of GPConv, enabling the fusion of information at both spatial and channel levels, thereby improving the average detection accuracy of the model. Secondly, ESPP utilizes four different scales of max-pooling operations, providing four different receptive fields, which helps the model better adapt to the recognition of multi-scale targets.

4.5. Detection Results for Different Categories of Targets

To validate the detection performance of YOLO-GP across different classes of objects, this study conducted multi-scale object detection experiments on both the dangerous behavior dataset and the Construction Site Safety Image Dataset. Subsequently, experiments were performed on different categories of objects within these datasets, and relevant metrics such as Precision (P) and Recall (R) were obtained. The final experimental results are presented in Table 5, where arrows in the table indicate the improvement status of the data.

Table 5. Comparison of detection performance between the baseline model and YOLO-GP model across different categories in the DBD and CSSID datasets.

		Baseline				**YOLO-GP**			
Dataset	**Category**	**Precision**	**Recall**	**mAP@.5**	**mAP@.5:.95**	**Precision**	**Recall**	**mAP@.5**	**mAP@.5:.95**
DBD	Helmet	89.9	84.6	88.4	44.9	93.5↑	85.0↑	89.9↑	45.6↑
	Head	92.3	90.8	94.0	61.2	91.9	91.2↑	94.4↑	61.4↑
	Phone	77.4	66.5	72.5	36.2	80.5↑	67.7↑	74.1↑	36.9↑
	Smoke	61.0	51.5	55.6	21.5	64.4↑	50.5	58.2↑	22.1↑
CSSID	Hardhat	96.4	60.8	73.6	42.5	91.7	64.6↑	74.3↑	43.0↑
	Mask	89.4	80.0	85.6	49.5	92.0↑	81.0↑	87.4↑	51.2↑
	No-hardhat	78.9	43.4	49.2	22.2	74.3	49.3↑	55.5↑	23.7↑
	No-mask	72.2	37.8	47.3	20.0	74.2↑	44.6↑	52.1↑	20.5↑
	No-safty vest	74.7	41.5	48.9	24.8	76.7↑	52.8↑	61.1↑	29.8↑
	Person	84.2	54.5	63.4	30.6	80.5	62.7↑	69.6↑	36.0↑
	Safty corn	88.9	70.5	77.0	36.6	74.7	75.0↑	76.1	38.1↑
	Safty vest	83.7	62.5	67.4	37.2	80.9	61.9	67.4	39.2↑
	Machinery	65.5	70.9	72.4	33.8	75.1↑	74.5↑	77.7↑	45.1↑
	Vehicle	72.7	23.8	29.5	17.2	93.1↑	32.0↑	41.4↑	23.3↑

From Table 5, it can be observed that the YOLO-GP model demonstrates varying degrees of performance improvement across different categories of objects. We analyze the reasons behind this phenomenon. Firstly, the DAM model structure effectively extracts richer gradient information, thereby enhancing the model's accuracy in object localization and its ability to discriminate between different targets. Secondly, the utilization of the ESPP module effectively enhances the model's feature extraction capability for multi-scale targets. Lastly, the CFE-Net effectively addresses the issue of insufficient channel correlation in the model, significantly enhancing the model's detection capability and accuracy in complex scenes. All three improvements significantly enhance the detection accuracy of different categories of objects in the dataset. Moreover, YOLO-GP achieves a certain degree of improvement across all categories, reflecting the model's versatility.

4.6. Visualization Results and Analysis

To effectively demonstrate the effectiveness of the algorithms in this study, two different datasets were visualized and analyzed in this study. A tuple assignment was used in this experiment to assign color values in RGB color space to each of the three variables detect_color, missing_color, and error_color. These color values are used in the experiment to represent the colors in different cases, respectively, to improve the accuracy and clarity of the visualization of the results. Specifically, detect_color represents the color (green) when a target is detected to identify the presence of the target; missing_color represents

the color (blue) when a target is not detected to help differentiate between undetected targets; and error_color represents the color (red) when an error situation occurs, which helps to quickly identify and locate the problem This helps to quickly identify and locate the problem. In this way, the state of the target in different situations during the detection process can be visualized more intuitively, which improves the interpretability of the results. The visualization of the baseline model and the YOLO-GP model for experiments on both datasets can be observed in the following figure.

1. Analysis of visualization results under DBD dataset.
   From Figure 11, it can be observed that the first and second columns of the visualized results exhibit characteristics such as high image grayscale and dim environments, which may adversely affect target detection. Grayscale images can result in blurred or lost target features and reduced contrast, thereby impacting the accuracy and stability of the algorithm. Dim environments may cause unclear target details and increased background noise, making it challenging for the target detection algorithm to correctly identify targets. The third and fourth columns of images both feature complex backgrounds and small targets. In such cases, the complex background may cause confusion between the target and the background, making it difficult for the algorithm to accurately locate and identify the target. Additionally, the presence of small targets may obscure the features of the target in the image, increasing the probability of false positives and false negatives in the detection algorithm. In these scenarios, the YOLO-GP model in the task of detecting dangerous behaviors among workers exhibits a significant reduction in the number of red and blue boxes in its visualizations compared to the baseline model. This reduction indicates a decrease in the probability of false positives and false negatives, thereby enhancing the reliability of detection.

Figure 11. Plots of visualization results of the baseline model and the YOLO-GP model under the DBD dataset. (**a**) Original image; (**b**) plot of visualization results of baseline model under DBD dataset; (**c**) plot of visualization results of YOLO-GP model under DBD dataset.

2. Analysis of visualization results under the CSSID dataset.
   In Figure 12, we observe characteristics such as cluttered backgrounds and significant differences in target scales. The cluttered background makes it challenging for the algorithm to distinguish targets from the surrounding environment, increasing the likelihood of false positives, especially when small targets are present. On the other hand, significant differences in target scales may lead to an imbalance in how the algorithm handles targets of different sizes, potentially resulting in detection errors or missing small targets. While any model has certain limitations, as shown in Figure 12, the YOLO-GP model also exhibits some false detections (red boxes) and missed targets (blue boxes). However, compared to the YOLOv8n model, the YOLO-GP model demonstrates better adaptability and robustness, achieving more accurate target detection and maintaining stable performance across different scales and background environments. Therefore, these visualized results further validate the superiority

of the YOLO-GP model in tackling complex backgrounds and multi-scale target detection tasks.

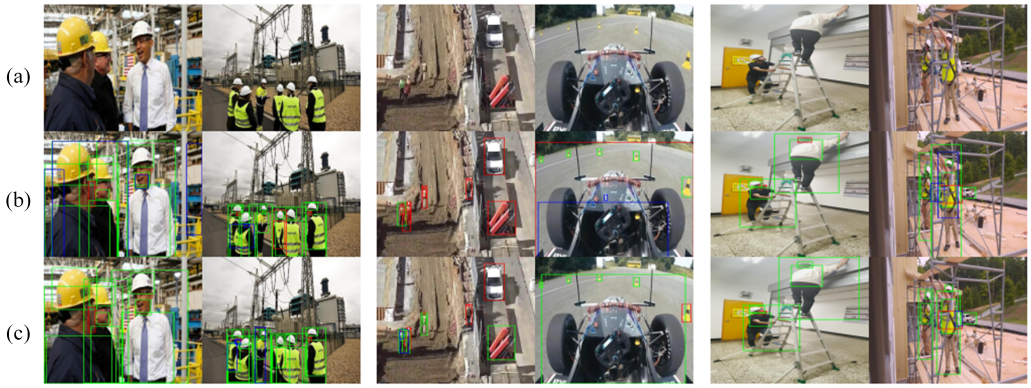

Figure 12. Plots of visualization results of baseline model and YOLO-GP model under CSSID dataset. (**a**) Original image; (**b**) visualization result plot of baseline model under CSSID dataset; (**c**) visualization result plot of YOLO-GP model under CSSID dataset.

4.7. Multi-Model Comparative Experiments

To validate the effectiveness of the proposed method, this study conducted comparative experiments on the dangerous behavior dataset with several mainstream object detection methods as well as the latest methods specifically designed for individual or multiple targets in this dataset. The experimental results are presented in Table 6, further confirming the feasibility and superiority of the improved model.

Table 6. Comparison of experimental results between YOLO-GP model and mainstream algorithmic models.

Dataset	Method	Precision	Recall	mAP@.5	mAP@.5:.95	GFLOPs	Inference Time/ms	Para/M
DBD	YOLOv3-tiny	77.8	71.2	74.5	40.8	19.1	0.7	12.17
	YOLOv5	76.8	73.3	74.8	40.0	4.2	9.7	1.78
	YOLOv6	75.7	70.8	74.2	43.8	13.1	0.5	4.50
	YOLOv7-tiny	76.5	71.8	74.6	39.4	13.2S	7.8	6.02
	YOLOv8n	80.1	73.3	77.6	44.9	8.9	9.9	3.16
	YOLO-CA	81.8	72.8	76.7	41.6	12.6	13.3	5.88
	YOLO-GP (Ours)	82.6	73.6	79.2	45.6	15.1	14.2	7.27
CSSID	YOLOv3-tiny	78.7	55.0	61.1	32.8	19.1	1.1	12.17
	YOLOv5	67.5	51.8	55.1	22.6	4.3	9.1	1.78
	YOLOv6	82.6	54.9	62.1	32.6	13.1	0.7	4.50
	YOLOv7-tiny	69.2	53.9	56.4	23.4	13.2S	7.0	6.02
	YOLOv8n	80.6	54.6	61.4	31.4	8.9	8.1	3.16
	YOLO-CA	75.8	60.0	65.1	28.8	12.6	11.9	5.88
	YOLO-GP (Ours)	81.3	59.8	66.3	35.0	15.1	11.7	7.27

From Table 6, it can be observed that compared to the YOLO-GP model, mainstream single-stage object detection algorithms such as YOLOv3-tiny [45], YOLOv5n [46], YOLOv6, YOLOv7-tiny, and YOLOv8n, although having smaller parameter counts and faster inference times, also demonstrate inferior detection performance on the two datasets used in this study. Specifically, these algorithms exhibit poorer performance in metrics such as mAP@.5 and mAP@.5:.95, failing to achieve the expected detection accuracy. Meanwhile, experiments comparing the latest helmet detection method, YOLO-CA [47], with the YOLO-GP detection model on the two datasets in this study show that although the former has certain computational efficiency compared to YOLO-GP, its detection accuracy under similar inference times still falls short of that of the YOLO-GP model. Specifically,

the YOLO-GP model achieves higher mAP@.5 and mAP@.5:.95 scores compared to the YOLO-CA model by 2.5%, 1.2%, and 4.0%, 6.2%, respectively, on the two datasets. This indicates the superiority of the YOLO-GP model in terms of detection accuracy, enabling more accurate detection of dangerous behaviors and effectively reducing the likelihood of dangerous occurrences.

5. Discussion

This study aims to propose a novel YOLO-GP-based model for dangerous behavior detection, enhancing the capability to detect dangerous behaviors of targets at various scales in complex environments, with tiny targets such as cigarettes and mobile phones. The model is built upon the YOLOv8n architecture and incorporates innovative modules such as GPConv, DAM, ESPP, and CFE-Net. Specifically, GPConv facilitates information exchange and feature fusion in the channel dimension, extracting rich feature representations. The DAM improves the Backbone and Neck structures, enhancing the richness of gradient flow information, thus addressing deficiencies in the accuracy of dangerous behavior target localization and small target recognition. The ESPP module significantly enhances dangerous behavior recognition by capturing and fusing multi-scale features. The CFE-Net module aids in better understanding the interactions between channels, improving the model's performance in detecting dangerous behaviors in complex scenes. Experimental results demonstrate that the YOLO-GP model improves the mAP@0.5:0.95 metric by 1.56% and 11.46% on dangerous behavior datasets and publicly available Construction Site Safety Image Datasets, respectively (as shown in Table 2), significantly outperforming the baseline model. Furthermore, compared to other single-stage pose estimation models, the YOLO-GP model achieves competitive performance on both datasets as shown in Table 6. The performance improvements validate the effectiveness of the proposed model in enhancing detection accuracy and robustness, indicating its promising prospects and significant value in practical applications.

However, despite the model's promising performance on the current dangerous behavior dataset and the publicly available Construction Site Safety Image Dataset, its performance in real-world scenarios may be influenced by additional factors such as lighting conditions, weather conditions, and background interference, which are often more complex and variable in real environments. Moreover, during the research implementation, data collection faced certain limitations. The datasets specifically designed for dangerous behavior detection in chemical enterprises are very limited, with most containing only one or two types of dangerous behaviors. Thus, a single behavior dataset may not fully cover all the variations and challenges present in real-world scenarios. Additionally, although YOLO-GP performs well on most metrics, its inference speed is not optimal. We attribute this to the introduction of the attention mechanism, which increases the complexity of the model structure, leading to a higher number of parameters and reduced model efficiency. While the computational efficiency of the model has not reached an optimal level, it still meets the requirements for most tasks.

Future work should not only focus on collecting more diverse and complex dangerous behavior datasets but also consider how to further improve the model's generalization ability across these diverse datasets. This may include introducing more data augmentation techniques to increase dataset diversity, designing more robust and transferable model architectures, and developing more intelligent and adaptive algorithms to tackle challenges in different scenarios. Additionally, efforts should be made to enhance model efficiency by simplifying the model structure, and reducing the number of parameters and computational load, thereby improving the model's speed and efficiency during inference. By reducing model complexity, we can better meet the demands of real-time behavior detection tasks, making the model more reliable and feasible for practical applications.

Author Contributions: Conceptualization, C.Y. and B.L.; methodology, C.Y.; validation, C.Y., B.L. and B.C.; data curation, Y.Z.; writing original draft preparation, C.Y.; writing—review and editing, B.C. and B.L.; visualization, C.Y.; supervision, Y.Z.; project administration, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Humanities and Social Sciences Project of the Ministry of Education of China under grant No. 22YJCZH014, the Natural Science Research Project of Jiangsu Provincial Universities under grant No. 22KJB520013, and the Natural Science Research Project of Huaiyin Institute of Technology under grant No. 22HGZ006 also supported this work.

Data Availability Statement: The datasets can be provided by the corresponding author upon reasonable request.

Acknowledgments: Thanks are due to Ling Wang and Shanshan Wang for their valuable discussion and the formatting of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lane, M.K.M.; Rudel, H.E.; Wilson, J.A.; Erythropel, H.C.; Backhaus, A.; Gilcher, E.B.; Ishii, M.; Jean, C.F.; Lin, F.; Muellers, T.D. ; et al. Green Chemistry as Just Chemistry. *Nat. Sustain.* **2023**, *6*, 502–512. [CrossRef]
2. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
3. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
4. 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]
5. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo Algorithm Developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
6. 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 14th European Conference on Computer Vision (ECCV 2016), Amsterdam, The Netherlands, 11–14 October 2016; Part I, pp. 21–37.
7. Bharati, P.; Pramanik, A. Deep learning techniques—R-CNN to mask R-CNN: A survey. In *Computational Intelligence in Pattern Recognition, Proceedings of the CIPR 2019, Howrah, India, 19–20 January 2019*; Springer: Singapore, 2020; pp. 657–668.
8. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [CrossRef] [PubMed]
9. Qian, X.; Wang, X.; Yang, S.; Lei, J. LFF-YOLO: A YOLO algorithm with lightweight feature fusion network for multi-scale defect detection. *IEEE Access* **2022**, *10*, 130339–130349. [CrossRef]
10. Ju, M.; Luo, J.; Wang, Z.; Luo, H. Adaptive feature fusion with attention mechanism for multi-scale target detection. *Neural Comput. Appl.* **2021**, *33*, 2769–2781. [CrossRef]
11. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
12. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
13. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FCANet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
14. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware global attention for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3186–3195.
15. Tan, M.; Pang, R.; Le Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
16. Rubaiyat, A.H.M.; Toma, T.T.; Kalantari-Khandani, M.; Rahman, S.A.; Chen, L.; Ye, Y.; Pan, C.S. Automatic Detection of Helmet Uses for Construction Safety. In Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), Omaha, NE, USA, 13–16 October 2016; pp. 135–142. [CrossRef]
17. Seshadri, K.; Juefei-Xu, F.; Pal, D.K.; Savvides, M.; Thor, C.P. Driver Cell Phone Usage Detection on Strategic Highway Research Program (SHRP2) Face View Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 35–43.
18. Wang, J. Driver Cell Phone Usage Detection Based on Semisupervised Support Vector Machine. Master's Thesis, Hunan University, Changsha, China, 2018.

19. Pan, G.; Yuan, Q.; Fan, C.; Qiao, H.; Wang, Z. Smoking Detection Algorithm Based on Mixture Gaussian Model and Frame Difference Method. *Comput. Eng. Des.* **2015**, *36*, 1290–1294.
20. Ai, B. Research on Indoor Cigarette Smoke Detection Algorithm Based on Video Surveillance. Master's Thesis, Yanshan University, Qinhuangdao, China, 2016.
21. Guo, Q.; Liu, N.; Wang, Z.; Sun, Y. Overview of Object Detection Algorithms Based on Deep Learning. *J. Detect. Control* **2023**, *45*, 10–20,26.
22. Dey, A.K.; Goel, B.; Chellappan, S. Context-Driven Detection of Distracted Driving Using Images from In-Car Cameras. *Internet Things* **2021**, *14*, 100380. [CrossRef]
23. Senyurek, V.Y.; Imtiaz, M.H.; Belsare, P.; Tiffany, S.; Sazonov, E. A Comparison of SVM and CNN-LSTM Based Approach for Detecting Smoke Inhalations from Respiratory Signal. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 3262–3265.
24. Han, G.; Li, Q. Rapid Smoking Detection Algorithm Based on Faster R-CNN. *J. Xi'An Univ. Posts Telecommun.* **2020**. [CrossRef]
25. Wang, Y. Research on Early Warning of Unsafe Behavior of Construction Workers Based on Convolutional Neural Network. Master's Thesis, Xi'an University of Architecture and Technology, Xi'an, China, 2021.
26. Chen, S.; Tang, W.; Ji, T.; Zhu, H.; Ouyang, Y.; Wang, W. Detection of Safety Helmet Wearing Based on Improved Faster R-CNN. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7.
27. Aboah, A.; Wang, B.; Bagci, U.; Adu-Gyamfi, Y. Real-Time Multi-Class Helmet Violation Detection Using Few-Shot Data Sampling Technique and Yolov8. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–20 June 2023; pp. 5349–5357.
28. Fan, X.; Wang, F.; Pang, S.; Wang, J.; Wang, W. Safety Helmet Wearing Detection Based on EfficientDet Algorithm. In Proceedings of the 2nd International Conference on Artificial Intelligence, Automation, and High-Performance Computing (AIAHPC 2022), Xi'an, China, 15–16 October 2022; SPIE: Bellingham, WA, USA, 2022; Volume 12348, pp. 302–309.
29. Yang, T.; Yang, J.; Meng, J. Driver's Illegal Driving Behavior Detection with SSD Approach. In Proceedings of the 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), Hangzhou, China, 13–15 November 2021; pp. 109–114.
30. Zhao, Z.; Zhao, H.; Ye, C.; Xu, X.; Hao, K.; Yan, H.; Zhang, L.; Xu, Y. FPN-D-Based Driver Smoking Behavior Detection Method. *IETE J. Res.* **2023**, *69*, 5497–5506. [CrossRef]
31. She, Y.; Zhang, X. Improved YOLOX Method for Small Target Smoking Detection Algorithm. In Proceedings of the International Conference on Cyber Security, Artificial Intelligence, and Digital Economy (CSAIDE 2023), Hangzhou, China, 14–15 September 2023; SPIE: Bellingham, WA, USA, 2023; Volume 12718, pp. 452–460.
32. Lei, Y.; Zhu, W.; Liao, H. Improved YOLOv8n Algorithm for Safety Helmet Detection in Complex Scenes. *Softw. Eng.* **2023**, *26*, 46–51. [CrossRef]
33. Zhang, T.; Qi, G.J.; Xiao, B.; Wang, J. Interleaved Group Convolutions. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4373–4382.
34. Hua, B.S.; Tran, M.K.; Yeung, S.K. Pointwise Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 984–993.
35. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
36. Zhang, X.; Zhang, Y.; Hu, M.; Ju, X. Insulator Defect Detection Based on YOLO and SPP-Net. In Proceedings of the 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Nanjing, China, 17–19 January 2020; pp. 403–407.
37. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
38. Roboflow Universe Projects. Construction Site Safety Dataset [Dataset]. Roboflow Universe. 2023. Available online: https://universe.roboflow.com/roboflow-universe-projects/construction-site-safety (accessed on 1 August 2023).
39. Bono, F.M.; Radicioni, L.; Cinquemani, S. A novel approach for quality control of automated production lines working under highly inconsistent conditions. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106149. [CrossRef]
40. Wan, D.; Lu, R.; Shen, S.; Xu, T.; Lang, X.; Ren, Z. Mixed Local Channel Attention for Object Detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106442. [CrossRef]
41. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
42. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–23 June 2023; pp. 7464–7475.
43. Weber, M.; Wang, H.; Qiao, S.; Xie, J.; Collins, M.D.; Zhu, Y.; Yuan, L.; Kim, D.; Yu, Q.; Cremers, D.; et al. Deeplab2: A Tensorflow Library for Deep Labeling. *arXiv* **2021**, arXiv:2106.09748.
44. Liu, S.; Huang, D. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.

45. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
46. Jocher, G.; Stoken, A.; Borovec, J.; NanoCode; Chaurasia, A.; Xie, T.; Liu, C.; Abhiram, V.; Laughing; tkianai ; et al. Ultralytics/yolov5: V5.0-YOLOv5-P6 1280 Models, AWS, Supervise.ly and YouTube Integrations. *Zenodo* **2021**.. [CrossRef]
47. Wu, X.; Qian, S.; Yang, M. Detection of Safety Helmet-Wearing Based on the YOLO-CA Model. *Comput. Mater. Contin.* **2023**, *77*, 3349–3366.