electronics

MDPI

Article

HR-YOLO: A Multi-Branch Network Model for Helmet Detection Combined with High-Resolution Network and YOLOv5

Yuanfeng Lian [1,2,*], Jing Li [2], Shaohua Dong [3,4] and Xingtao Li [5]

1 Beijing Key Lab of Petroleum Data Mining, Beijing 102249, China
2 Department of Computer Science and Technology, China University of Petroleum, Beijing 102249, China; 2022216025@student.cup.edu.cn
3 College of Safety and Ocean Engineering, China University of Petroleum, Beijing 102249, China; shdong@cup.edu.cn
4 Key Laboratory of Oil and Gas Safety and Emergency Technology, Ministry of Emergency Management, Beijing 102249, China
5 China National Oil and Gas Exploration and Development Co., Ltd., Beijing 102100, China; lixingtao@cnpcint.com
* Correspondence: lianyuanfeng@cup.edu.cn

Abstract: Automatic detection of safety helmet wearing is significant in ensuring safe production. However, the accuracy of safety helmet detection can be challenged by various factors, such as complex environments, poor lighting conditions and small-sized targets. This paper presents a novel and efficient deep learning framework named High-Resolution You Only Look Once (HR-YOLO) for safety helmet wearing detection. The proposed framework synthesizes safety helmet wearing information from the features of helmet objects and human pose. HR-YOLO can use features from two branches to make the bounding box of suppression predictions more accurate for small targets. Then, to further improve the iterative efficiency and accuracy of the model, we design an optimized residual network structure by using Optimized Powered Stochastic Gradient Descent (OP-SGD). Moreover, a Laplace-Aware Attention Model (LAAM) is designed to make the YOLOv5 decoder pay more attention to the feature information from human pose and suppress interference from irrelevant features, which enhances network representation. Finally, non-maximum suppression voting (PA-NMS voting) is proposed to improve detection accuracy for occluded targets, using pose information to constrain the confidence of bounding boxes and select optimal bounding boxes through a modified voting process. Experimental results demonstrate that the presented safety helmet detection network outperforms other approaches and has practical value in application scenarios. Compared with the other algorithms, the proposed algorithm improves the precision, recall and mAP by 7.27%, 5.46% and 7.3%, on average, respectively.

Keywords: deep learning; safety helmet; human pose; optimized residual network; attention module; NMS

check for updates

Citation: Lian, Y.; Li, J.; Dong, S.; Li, X. HR-YOLO: A Multi-Branch Network Model for Helmet Detection Combined with High-Resolution Network and YOLOv5. *Electronics* **2024**, *13*, 2271. https://doi.org/10.3390/electronics13122271

Academic Editors: Aryya Gangopadhyay and Phill Kyu Rhee

Received: 15 April 2024
Revised: 12 May 2024
Accepted: 5 June 2024
Published: 10 June 2024

Copyright: © 2024 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Helmet-wearing detection is an important application in computer vision that has been widely used in various fields, such as traffic surveillance [1–3], smart cities [4,5], construction safety [6–8], etc. The detection of safety helmet wearing can be summarized as identifying multi-scale targets that are easily occluded in complex environments. Although some breakthroughs have been made with respect to this problem [9,10], it is still challenging.

Traditional methods for helmet detection [11–15] have lower accuracy and poorer robustness in their results due to the need for manual inspection, which is time-consuming and prone to human mistakes. With significant advancements in deep learning within the

field of object detection, this technique has been further used to improve helmet detection algorithms [16–21]. These methods greatly improve the accuracy and efficiency of detection and can carry out real-time monitoring. Helmet recognition methods using deep learning can be divided into the following two categories: two-stage methods [22–25] and one-stage methods [26–34]. Two-stage methods extract features, generate candidate regions and use classifiers for classification and regression. Currently, the Region-based Convolitional Neural Network (R-CNN) series generates candidate regions using a region proposal (RP) algorithm and eventually uses a classifier for feature classification. However, faced with redundant parameters, high computational complexity and slow inference speeds, those methods cannot meet the requirement of real-time detection. The one-stage methods use an end-to-end strategy to detect and classify the target position of the image. The Single Shot MultiBox Detector (SSD) series incorporates both regression and an anchor mechanism to predict on multi-scale feature maps, encountering difficulty in small-target detection owing to its strategy of not merging features. The YOLO series judges the target's category and position at the same time by transforming the object detection problem into a regression problem. The advantage of two-stage methods is that they can effectively improve the detection accuracy, but it is difficult to achieve real-time detection. Although the one-stage make it difficult to achieve model convergence during intensive sampling, with poor performance for small targets, they are more suitable for the real-time detection in industrial production scenarios. For safety helmet detection, the You Only Look Once (YOLO) series has advantages in terms of a faster detection speed, better real-time performance and easier model deployment.

Although there has been significant progress in safety helmet detection, the above-mentioned methods ignore human pose information. In complex industrial scenarios, the detection of safety helmets requires the fusion of human pose information, which provides additional clues for safety helmet wearing identification. In response to the aforementioned issue, we propose a new deep learning method called High-Resolution You Only Look Once (HR-YOLO) based on a high-resolution network and YOLOv5 to enhance the feature representation capabilities. To provide efficient feature extraction and description, we design an optimized network structure called Optimized Powered Stochastic Gradient Descent (OP-SGD) for the construction of HR-YOLO. A novel self-attention model, namely the Laplace-aware attention model (LAAM), is then used to extract and fuse the features from the object detection branch (ODB) and the pose detection branch (PDB). More specifically, to address the issue of the limited reliability of human pose for helmet wearing detection, we design pose-assisted non-maximum suppression voting (PA-NMS Voting) to select features with high reliability.

Our main contributions can be summarized as follows:

- We propose a novel HR-YOLO to achieve high-quality feature fusion between helmet object detection and human pose estimation in which the pose detection branch (PDB) can effectively exchange the input image with the required human pose information and the object detection branch (ODB) can extract helmet features from backbone features. In addition, we design PDB loss and ODB loss to construct the HR-YOLO loss function.
- An optimized network structure, OP-SGD, is proposed to optimize the structure, enhance the expressive ability of the network and accelerate the speed of convergence.
- We design a Laplace-aware attention model (LAAM) for feature enhancement. LAAM can highlight the local neighborhood that contains fine-grained structural information and make HR-YOLO pay more attention to the pose feature, which improves the detection accuracy for occluded and small objects. We also propose PA-NMS by using human pose information constraints to modify the selection method of NMS and further propose non-maximum suppression voting (PA-NMS voting) to select the optimal bounding box, which improves the accuracy of localization.
- Our method outperforms the YOLOv5 methods in experiments of safety helmet wearing detection on the GDUT-HWD and SHWD datasets. The ablation study shows a significant improvement when integrating OP-SGD and LAAM in HR-YOLO.

2. Related Works

2.1. Helmet Detection

Helmet detection algorithms are divided into traditional machine learning-based and deep learning-based methods. Traditional machine learning methods used steps that include background subtraction, human detection and helmet recognition, utilizing manually selected features or statistical features [35,36]. However, the traditional methods, relying on hand-engineered features, cannot achieve good real-time speed and detection accuracy due to their multi-stage operation. Since its advent, deep learning has been used with computer vision algorithms, which can be divided into two detection types, namely two-stage and one-stage algorithms. The former is represented by the region-based convolutional neural network (R-CNN) series [22–25]. The latter includes the YOLO series [26,30–34], the SSD series [26–29], RetinaNet [37], etc. These technologies have been widely used for safety helmet detection and have accomplished many achievements. Wang et al. [17] proposed the Faster R-CNN algorithm to inspect the wearing of safety helmets. Long et al. [38] presented a deep learning approach for accurate safety helmet wearing detection in employing an SSD. To date, several YOLO-based models have achieved an accuracy level of approximately 90%, which satisfies the demands of real-time detection in construction site scenarios. Numerous researchers have studied safety helmet wearing detection algorithms based on the YOLO series. Jamtsho et al. [39] presented the real-time detection of LP for non-helmeted motorcyclist using a YOLO real-time object detector. Wu et al. [40] utilized an improved model based on YOLOv3 to detect helmet wearing. Zhou et al. [41] proposed a safety helmet detection method using the YOLOv5 model. However, those methods only focus on the target features of safety helmets and ignore the human pose features, which can provide position information about helmets. Inspired by [42,43], we designed High-Resolution YOLO (HR-YOLO) combined with HRNet and YOLOv5 to take multi-branch information from helmet object detection and human pose estimation, which improves recognition accuracy in the cases of target occlusion, insufficient light, complex backgrounds and other problems.

2.2. Optimized Network Structure Design

Extensive work has been conducted in the field of neural network structural optimization. In the early stages of neural network development, evolutionary algorithms were commonly employed to find optimal architectures and weights [44]. Matias et al. [45] utilized the genetically optimized extreme learning machine (GO-ELM) to optimize the network architecture. Some researchers [46,47] have employed an adaptive strategy to progressively expand the network structure layer by layer from a small network guided by specific principles. However, the achieved results failed to clearly show where the connections should be established in the network architecture. In Ref. [48], it was demonstrated that the design of a neural network structure can be motivated by faster optimization algorithms. Furthermore, Lu et al. [49] bridged deep neural network design with numerical differential equations and interpreted different CNN models with residual blocks and special discretization schemes. The gradient descent (GD) algorithm [50] is one of the most widely employed optimization methods and serves as the foundation for various other optimization algorithms. To enhance the iteration speed and improve the accuracy of the model, we designed an OP-SGD method incorporating the pbSGD [51] algorithm, which can optimize the propagation structure of the residual network to enhance the input descriptors of features.

2.3. Attention Mechanism

In the field of computer vision, attention mechanisms have gained considerable interest due to their ability to efficiently focus on the representations of Regions of Interest (ROIs) in images or videos. Researchers have proposed many attention models for objection detection, such as class activation mapping (CAM) [52], a stereo attention module (SAM) [53], a convolutional block attention module (CBAM) [54] and squeeze and excitation (SE) [55].

Furthermore, Anwar et al. [56] presented the densely residual Laplacian network (DRLN), which uses Laplacian attention to model the crucial features. Due to the great success of attention modules, they have been used in helmet detection. Han et al. [57] proposed a novel SSD algorithm for safety helmet wearing detection that uses a cross-layer attention mechanism to further refine the feature information of the object region. Chen et al. [58] used a YOLOv5 image classifier combined with residual transformer spatial attention to detect riders' helmet wearing. Tai et al. [7] employed an attention mechanism to improve the model's generalization ability, which aligned with practical application requirements. Inspired by these methods, our paper presents the a Laplace-aware attention module to improve the accuracy and robustness of helmet detection.

3. Proposed Method

3.1. HR-YOLO Network

In order to accomplish multi-scale hard-hat target detection, a multi-branch parallel fusion network called HR-YOLO was constructed, as depicted in Figure 1. The original image, with a size of 640 × 640 × 3, first passes through the backbone part for extraction of preliminary feature information. These features pass to the ODB, which further extracts and filters out features for helmet object detection by using the LAAM. Simultaneously, the PDB extracts the human pose features from preliminary features and determines the head-area information. Finally, PA-NMS voting uses helmet object features and head-area information to select bounding boxes with higher confidence and reliability. Through these four improvements, HR-YOLO effectively recognizes small objects and performs compliance detection for safety helmet wearing.

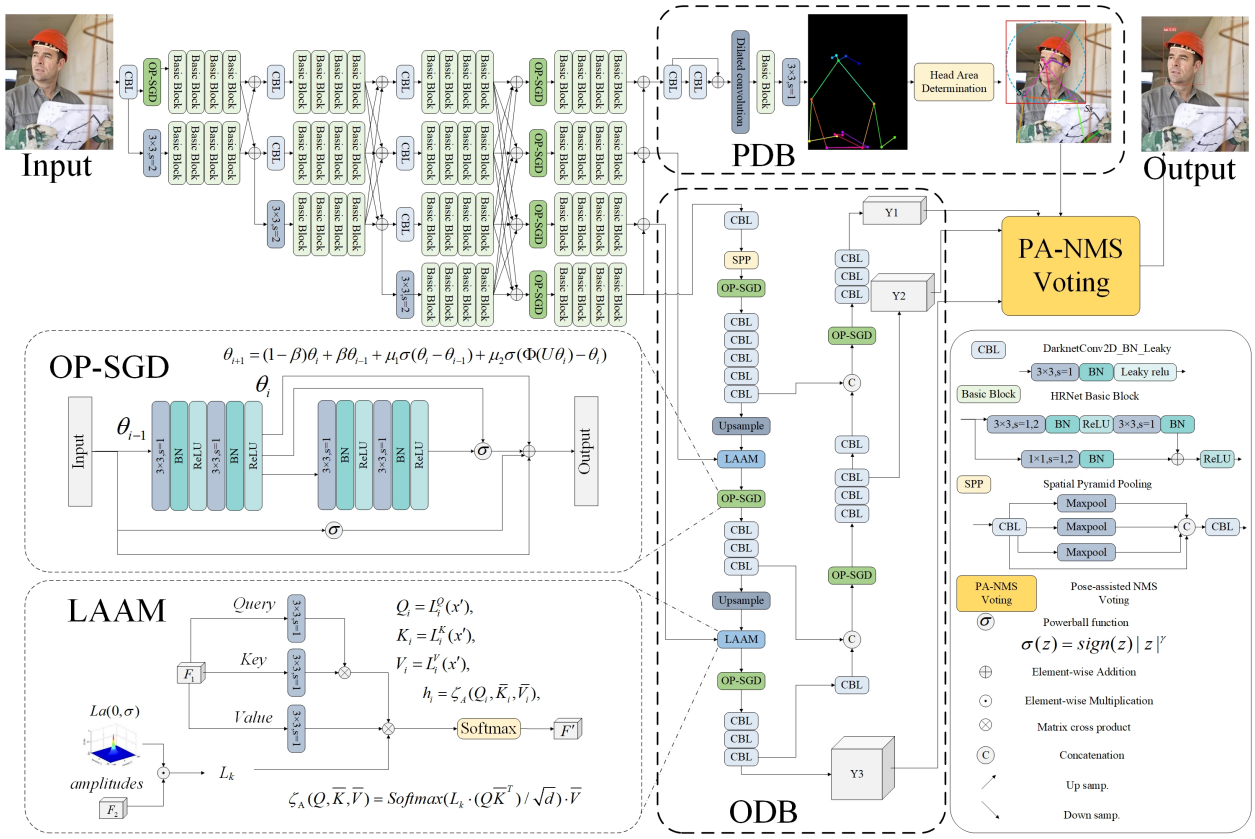


Figure 1. HR-YOLO network architecture.

The detection of the head region plays a crucial role in providing the necessary location information for compliance discrimination and enhancing the target feature expression capability of the detection branch. The PDB utilizes HRNet as the neural network to identify

key points of the human body, enabling the determination of the target's position through the creation of head-region judgment rules. HRNet effectively captures 18 key points of the human body, as illustrated in Figure 2a. This accurate identification of key points aids in locating the head region, which is essential for safety helmet compliance detection.

Definition 1. *(Head area of upper body): We assume that there is* $\{x_i, y_i, c_i\} \in \mathrm{U}$ *and* U *is not empty in the key point set of the PDB.*

$$D_{K_1,K_2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

$$D_{K_1,K_3} = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} \tag{2}$$

$$radius = \min\{D_{K_1,K_2}, D_{K_1,K_3}\} \tag{3}$$

$$H_{area} = Rect(Circle(K_1, radius)) \tag{4}$$

where $K_1(x_1, y_1)$, $K_2(x_2, y_2)$ *and* $K_3(x_3, y_3)$ *are the key points for the nose, left shoulder and right shoulder, respectively.* D_{K_1,K_2} *is the Euclidean distance between the nose key point and the left-shoulder key point, and* D_{K_1,K_3} *is the Euclidean distance between the nose key point and the right-shoulder key point.* $Circle(K_1, radius)$ *is a circle with* $K_1(x_1, y_1)$ *as the center and radius as the radius.* $Rect()$ *is the outer square of* $Circle$. H_{area} *is the head area of the object. Each key point and positioning area is shown in Figure 2b.*

Definition 2. *(Head area of face): We assume that there is* $\{x_i, y_i, c_i\} \in \mathrm{U}, i = 1, 2, \cdots, 5$ *and* U *is not empty in the key point set of the PDB.*

$$Angles[Pitch, Yaw, Roll] = F(\{x_i, y_i, c_i\}) \tag{5}$$

where $\{x_i, y_i\}$ *represents the coordinates of the key points for the left eye, right eye, left ear, right ear and nose, respectively.* $\{c_i\}$ *represents the confidence level of each point.* $Pitch$, Yaw *and* $Roll$ *are the yaw angle, pitch angle and roll angle, respectively.* F *is a head-pose mapping function defined as the head-pose estimation network in [59]. The head-pose estimation is shown in Figure 2c.*

(a) (b) (c) (d)

Figure 2. Head area determination rules. (**a**) 18 key points of the human body. Points 0 to 17 represent nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, left ankle, right eye, left eye, right ear, and left ear, respectively. (**b**) The result of the head area calculated by the upper body points.(**c**) The result of head-pose estimation. (**d**) The result of the head area calculated by the whole body points.

Definition 3. *(Head area of whole body): We assume that there is* $K_1(x_1, y_1) \in \mathrm{U}$ *and* U *is not empty in the key point set of the PDB.* $K_4(x_4, y_4)$, $K_5(x_5, y_5)$ *and* $K_6(x_6, y_6) \in \mathrm{U}$, *and* $K_0(x_0, y_0)$ *is the regional center the following key point:*

$$D_{K_4,K_5} = \sqrt{(x_4 - x_5)^2 + (y_4 - y_5)^2} \tag{6}$$

$$D_{K_4,K_6} = \sqrt{(x_4 - x_6)^2 + (y_4 - y_6)^2} \tag{7}$$

where $K_4(x_4, y_4)$, $K_5(x_5, y_5)$ *and* $K_6(x_6, y_6)$ *are the key point coordinates of the neck, left hip and right hip, respectively.* D_{K_4,K_5} and D_{K_4,K_6} *are the Euclidean distances between the neck key point and the left-hip key point and the right-hip key point, respectively. The distance between the key point of the neck* ($K_4(x_4, y_4)$) *and the key point of the center of the head object area* ($K_0(x_0, y_0)$) *is defined as follows*:

$$\overrightarrow{K_4, K_0} = \frac{\overrightarrow{K_4K_5} + \overrightarrow{K_4K_6}}{\left|\overrightarrow{K_4K_5} + \overrightarrow{K_4K_6}\right|} \tag{8}$$

$$radius = D_{K_4,K_0} = \lambda \min\left\{D_{K_4,K_5}, D_{K_4,K_6}\right\} \tag{9}$$

$$H_{area} = Rect(Circle(K_0, radius)) \tag{10}$$

where D_{K_4,K_0} *is the Euclidean distance between the key point of the neck and* $K_0(x_0, y_0)$. $\overrightarrow{K_4, K_0}$ *is the unit vector.* λ *is the proportionality factor. Each key point and positioning area is shown in Figure 2d.*

3.2. PA-NMS Voting

We proposed pose-assisted non-maximal suppression (PA-NMS) to replace traditional NMS in the bounding-box vote process [60], obtaining PA-NMS voting. As shown in Algorithm 1, a non-maximal suppression PA-NMS vote is proposed to select bounding boxes with pose information constraints.

Algorithm 1 Algorithm of PA-NMS

Input: initial target detection box $B = b_1, b_2 \cdots b_N$, target confidence $S = s_1, s_2 \cdots s_N$, key point detection box $P = p_1, p_2 \cdots p_M$, detection box decision coefficient λ, NMS threshold N_t, PA-NMS threshold N_p

Output: final detection box $D = d_1, d_2 \cdots d_N$, detection scores S, safety helmet wearing decision coefficient $X = x_1, x_2 \cdots x_N$

$D \leftarrow \{\}, X \leftarrow \{false\}$
while $B \neq empty$ **do**
  for b_i in B **do**
    for p_j in P **do**
      if $IoU(b_i, p_i) \geq N_p$ **then**
        $s_i \leftarrow (1 - \lambda) \cdot s_i + \lambda$
        $x_i \leftarrow true$
      end if
    end for
    if $x_i = false$ **then**
      $s_i \leftarrow \lambda \cdot s_i$
    end if
  end for
  $m \leftarrow \arg\max S$
  $M \leftarrow b_m, x_i \leftarrow m$
  $D \leftarrow D \cup M, B \leftarrow B - M$
  for b_i in B **do**
    if $b_i \in B$ and $IoU(M, b_i) \geq N_t$ **then**
      $s_i \leftarrow s_i \cdot (1 - IoU(M, b_i))$
    end if
  end for
end while
return D, S, X

We find the maximum confidence ($S_{\max}$) in confidence set S and judge by calculating the intersection over union (IoU) of the border corresponding to $S_{\max}$ and other confidence values. When $IoU(b_m, p_i) \geq N_p$, the head area of the pose key point positioning can be regarded as a judgment constraint on the wearing state of safety helmets.

The PA-NMS is designed to improve the ability to detect the behavior of incorrect helmet wearing by performing IoU calculation using pose information and bounding boxes. After using PA-NMS, we obtained the desired results ($Y = \{(S_j, B_j)\}, j \in N_+$). The voting Mechanism regards each prediction result as an independent score and votes on all the prediction results together in a weighted way to find the optimal bounding box. However, to address the issue of excessively low weights assigned to certain boxes during the voting process, we make modifications to the traditional voting process. The vote progress is defined as follows:

$$\omega_j = \max(0.01, s_j) \tag{11}$$

$$B_i' = \frac{\sum_{j:B_j \in B} \omega_j \cdot B_j}{\sum_{j:B_j \in B} \omega_j} \tag{12}$$

where s_j represents the detection scores from PA-NMS, and B corresponds to the initial target detection boxes.

3.3. OP-SGD

In the context of feed-forward networks, the issues of gradient disappearance and network degradation can be effectively addressed by optimizing the residual structure, improving network performance [48]. To enhance the expressive ability of features within the residual block, we propose a novel optimized variant of the Powered Stochastic Gradient Descent (pbSGD) algorithm [51] called OP-SGD. OP-SGD is designed to fuse the knowledge modeling capability of the optimization strategy and the adaptive learning ability of the deep learning method to enhance the interpretability of optimized residual network infrastructure, establishing a connection between the pdSGD optimization algorithm and its corresponding neural architecture. The derivation process for OP-SGD can be expressed as follows, utilizing the pbSGD formula:

$$x_{t+1} = x_t - \alpha\sigma(g_t) \tag{13}$$

$$x_{t+1} = (1-\beta)x_t + \beta x_{t-1} - \alpha\beta\sigma(g_{t-1}) - \alpha\sigma(g_t) \tag{14}$$

where x_i is an arbitrary point, α is the learning rate at step t and g_t is stochastic gradient. $\sigma = sign(z)|z|^\gamma$ is named the Powerball function. β is a parameter used to control the influence of x_{t-1} on x_{t+1}.

We assume that during the propagation of the neural network, the transmission of the signal from the first layer to the last layer is represented as follows:

$$\theta_{i+1} = \Phi(U_i\theta_i) \tag{15}$$

where θ_{k+1} denotes the characteristics of layer $k+1$ in the network. Φ is an activation function. Assume that U is a symmetric positive definite matrix ($V = \sqrt{U}$) and $\xi = V\theta$; then, for a nonlinear activation function ($\Phi(\xi)$), there exists a function ($\Psi(\xi)$) such that when $\Psi'(\xi) = \Phi(\xi)$,

$$\nabla \sum_i \Psi\left(V_j^T\xi\right) = U\Phi\left(U^T\theta\right) = U\Phi(U\theta) \tag{16}$$

The object function ($f(\xi)$) is defined as follows:

$$f(\xi) = \frac{\left\|\xi^2\right\|}{2} - \sum_i \Psi\left(V_j^T\xi\right) \tag{17}$$

where V_i is the ith column of V.

We derive the derivative on both sides of the function ($f(\xi)$) to obtain the following:

$$\nabla f(\xi_i) = \xi_i - V\Phi(V\xi_i) \tag{18}$$

Furthermore, Equation (14) can be expressed as follows:

$$\xi_{i+1} = (1-\beta)\xi_i + \beta\xi_{i-1} - \mu_1\sigma(\nabla f(\xi_i)) - \mu_2\sigma(\nabla(f(\xi_2)) \tag{19}$$

We obtain the following according to $\theta = V^{-1}\xi$:

$$\theta_{i+1} = (1-\beta)\theta_i + \beta\theta_{i-1} + \mu_1\sigma(\theta_i - \theta_{i-1}) + \mu_2\sigma(\Phi(U\theta_i) - \theta_i) \tag{20}$$

where $\Phi(U\theta_i)$ denotes the ith layer of the feed-forward network.

According to Equation (20), a neural network structure, OP-SGD, is designed as an optimized residual network structure with two shortcuts. The network structure inspired by the structure corresponding to this formula is shown in Figure 1.

3.4. Laplace-Aware Attention Module

Pixel-wise matching is susceptible to poor feature quality. However, complex background and illumination changes reduce feature discrimination. To avoid ambiguities, we propose a Laplace-aware attention module (LAAM) to obtain the fine-grained structural information and locally discriminative representations for feature matching, as shown in Figure 1. Specifically, given base feature x, it is formulated as follows:

$$\hat{x} = LAA(LN(x)) + x, \tag{21}$$

$$y = FFN(LN(\hat{x})) + \hat{x}, \tag{22}$$

where $LN(\cdot)$ is general layer normalization and $FFN(\cdot)$ is the feed-forward network in the block. $LAA(\cdot)$ is Laplace-aware attention, the core component of our LAAM.

We formulate LAA as a task-specific local operation, which not only avoids a misleading global context but also reduces the complexity of attention computation. Therefore, it can be formulated as follows:

$$Q_i = L_i^Q(x'), K_i = L_i^K(x'), V_i = L_i^V(x'), \tag{23}$$

$$h_i = \zeta_A(Q_i, \bar{K}_i, \bar{V}_i), \tag{24}$$

$$H = Concat(h_1, h_2, ..., h_n), \tag{25}$$

where $x' = LN(x)$. $L_i^Q(\cdot)$ and $L_i^V(\cdot)$ denote linear projections for the ith head. $\zeta_A(\cdot)$ is the Laplace attention function, which takes query feature Q_i. The regional features of key $\bar{K}_i$ and value $\bar{V}_i$ are defined as follows:

$$\zeta_A(Q, \bar{K}, \bar{V}) = Softmax(L_k \cdot (Q\bar{K}^T)/\sqrt{d}).\bar{V} \tag{26}$$

where L_k is a learnable Laplace kernel with dimensions of$k \times k$, which can be updated by adding a learnable amplitude matrix (A) during model training. In the inference process, L_k can reorganize the weights of attentive feature aggregation.

Assume the neighborhood of a pixel at point p is $N(p)$; then, the attention on a single pixel can be defined as follows:

$$h_{(p)} = Softmax(L_k \cdot (Q_{(p)}\bar{K})^T_{N(p)}/\sqrt{d}) \cdot \bar{V}_{N(p)} \tag{27}$$

Note that the operating range is salable with the varying region of $N(p)$. For instance, it can be extended to all pixels (i.e., $N(p)$ is equivalent to the image size), leading to global self-attention in a Laplace-aware manner.

3.5. Compliance Reasoning Decision Algorithm

The compliance reasoning decision process for helmet wearing is illustrated in Algorithm 2. First, we generate the coordinates of the key points for the human body by using the human posture detection branch. Next, we determine the region of interest based on the head-region determination rule. Finally, we utilize the PA-NMS algorithm to identify the target anchor frame ratio within the target area of the two branches. This enables us to assess compliance with safety helmet wearing regulations.

Algorithm 2 Compliance Reasoning Decision Algorithm

Input: oil and gas field image I, threshold value τ
Output: probability of safety helmet test results and wearing P

Images I are input into PDB to obtain 18 key point coordinates $k_i, i = 1, 2 \cdots 18$
Images I are input into ODB, obtain regression box $B = \{b_1, b_2 \cdots b_N\}$, height of box $H = \{h_1, h_2 \cdots h_N\}$, the distance between key point of nose k_0 and top of box $D = \{d_1, d_2 \cdots d_N\}$, and degree of confidence $CLS = \{cls_1, cls_2 \cdots cls_N\}$
Select k_i to determine head area H_{area} by the rule of head area assessment
Obtain head pose $Angles[Pitch, Yaw, Roll]$ by the rule of head pose estimation
Use PA-NMS model in Algorithm 1 to generate final detection box D and confidence score S
if $\tau < s_i < 1$ and $cls_i = hat$ **then**
    $p_i \leftarrow true$
end if
if $0 < s_i < \tau$ and $cls_i = hat$ and $d_i / \cos(roll) > 2h/3$ **then**
    $p_i \leftarrow true$
end if
if $\tau < s_i < 1$ and $cls_i = person$ **then**
    $p_i \leftarrow false$
end if
if $0 < s_i < \tau$ and $cls_i = person$ and $k_0 \in h_i$ **then**
    $p_i \leftarrow false$
end if

3.6. Loss Functions

HR-YOLO has a multi-branch structure, and the overall loss function (L) is the weighted sum of multiple loss functions. As a result, the overall loss function is defined as the weighted sum of multiple loss functions. L can be expressed as follows:

$$L = \lambda_1 L_{ODB} + \lambda_2 L_{PDB} \tag{28}$$

where L_{ODB} and L_{PDB} correspond to the loss functions of ODB and PDB, respectively. λ_1 and λ_2 are the coefficients of each loss function; we set λ_1 as 0.5 and λ_2 as 0.5.

The L_{ODB} loss function in the detection-branch training process is defined as follows:

$$L_{ODB} = L_{cls} + L_{reg} + L_{conf} \tag{29}$$

where L_{cls}, L_{reg} and L_{conf} are the classification loss, positioning loss and confidence loss of the target regression, respectively.

The L_{PDB} loss function in each stage of the training process of the pose-detection branch is defined as follows:

$$L_{PDB} = \sum_{t=1}^{T} (L_{PCM} + L_{PAF}) \tag{30}$$

where L_{PCM} is the key point position prediction loss, and L_{conf} is the joint vector loss. We define them as follows:

$$L_{PCM} = \sum_{j=1}^{J} \sum_{p} W(p) \cdot \left\| S_j^t(p) - S_j^*(p) \right\|_2^2 \tag{31}$$

$$L_{PAF} = \sum_{c=1}^{C} \sum_{p} W(p) \cdot \left\| L_c^t(p) - L_c^*(p) \right\|_2^2 \tag{32}$$

where s_j^* is the confidence map, L_c^* represents the position vector and $W(p)$ is the width in p. c and j are the position vector number and the key point number, respectively. When the image is missing at position p, $W(p) = 0$.

4. Experiments and Implementation

4.1. Experimental Environment

The experimental platform uses a Nvidia GTX2080TI-11G graphics card, Taiwan, China (GPU processing unit), and the deep learning framework is Pytorch 1.7.1. The experimental datasets include the GDUT-HWD, SHWD and self-made datasets. The embedded test platform carried by the quadruped robot is Nvidia Jetson TX2,256 CUDA cores, Taiwan, China with 8 GB memory and 32 GB storage space. After the model is trained on the experimental platform, it is transplanted to the embedded test platform. In the posture-detection branch, the ratio of the distance from the eye to the shoulder to the distance from the shoulder to the hip is $\lambda = 0.35$.

4.2. Performance Metric

We use *precision*, *recall*, mAP and IoU as evaluation metrics in order to evaluate the model's accuracy in safety helmet detection. The calculation formula is expressed as follows:

$$precision = \frac{TP}{TP + FP} \tag{33}$$

$$recall = \frac{TP}{TP + FN} \tag{34}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{35}$$

$$IoU = \frac{DR \cap GT}{DR \cup GT} \tag{36}$$

where TP indicated a positive model prediction with positive values, FP is a positive model prediction with negative values and FN is an incorrect negative prediction by the model. DR is the safety helmet area framed after detection, while GT is a standard value. AP is the average accuracy of each class, and C is the class number.

4.3. Results

To conduct a quantitative analysis of the detection performance of the proposed method compared with other methods, we present the experimental results of metrics on the GDUT-HWD dataset in Table 1. The comparison includes SSD [26], R-SSD [27], Faster RCNN [24], YOLOv3 [31], YOLOv3-tiny [31], YOLOv4 [32], YOLOv5s, YOLOv7s [34] and YOLOv8s. Taking the mean average precision (mAP) evaluation metric as an example, the proposed method achieves higher detection accuracy in five categories compared with the other methods. Similarly, when evaluating precision and F1 scores, the proposed method demonstrates improved detection accuracy across different categories compared with the alternative methods.

As shown in Table 2, on the SHWD dataset, when the IoU threshold is set to 0.5, the detection accuracy of the proposed method reaches 96.1%. The accuracy for the hat

category is 97.4%, and accuracy for the person category is 94.9%. When the IoU threshold is between 0.5 and 0.95, the mAP reaches 65.4%, which is 4.3% higher than that of YOLOv5s.

As shown in Figure 3, HR-YOLO can detect targets well in images with dense crowds, complex backgrounds and occluded targets. Similarly, the proposed algorithm can still detect targets accurately and robustly in images with insufficient light at night.

Table 1. Experimental results comparing different algorithmic models on the GDUT-HWD dataset.

Method	P	R	F1	AP@0.5					mAP@0.5	mAP@0.5:0.95
				White	Blue	Red	Yellow	None		
SSD	79.3	64.4	71.2	59.6	64.0	71.8	80.6	78.6	70.9	42.9
R-SSD	80.9	78.7	79.8	85.6	79.1	81.3	88.0	83.4	83.5	56.5
Faster RCNN	85.0	82.7	83.8	89.9	83.1	85.5	92.4	87.7	87.8	59.4
YOLOv3	87.6	85.3	86.4	92.7	85.7	88.1	92.3	88.5	89.5	60.2
YOLOv3-tiny	88.1	71.8	79.1	66.2	71.1	79.8	89.5	87.3	78.8	47.7
YOLOv4	91.4	81.1	86.2	86.1	95.7	94.2	92	83.4	90.3	61.3
YOLOv5s	91.1	87.5	89.3	87.9	90.1	93.6	92.6	82.3	89.3	54.3
Yolov7s	91.8	80.1	85.5	84.9	94.4	92.9	90.7	82.3	89.1	62.6
Yolov8s	88.8	80.8	84.3	83.6	85.7	89.1	88.1	78.3	85	53.4
Ours	92.8	86.7	89.6	93.6	91.6	93.3	95.6	84.9	91.8	62.9

Table 2. Experimental results comparing different algorithmic models on the SHWD dataset.

Method	P	R	F1	Class		mAP@0.5	mAP@0.5:0.95
				Hat	Person		
SSD	81.6	75.9	78.6	83.8	73.1	78.5	46.5
R-SSD	84.9	86.9	85.9	86.9	89.1	88.0	58.7
Faster RCNN	91.8	91.3	91.5	91.3	93.6	92.5	61.7
YOLOv3	90.7	91.7	91.2	91.8	93.7	92.8	60.1
YOLOv3-tiny	86.7	75.8	80.9	84.6	74.3	79.4	46.2
YOLOv4	90.4	88.7	89.5	95.3	93.9	94.6	59.3
YOLOv5s	93.5	90.3	91.8	96.1	93.9	94.7	61.1
YOLOv7s	91.8	87.6	89.6	96.1	90.6	93.3	60.6
YOLOv8s	92.3	83.4	87.6	85.5	95.7	90.2	60
Ours	94.7	94.3	94.4	97.4	94.9	96.2	65.4

(a)YOLOv4 (b)YOLOv5s (c)YOLOv8s (d)Ours

Figure 3. Some examples of detection results on the SHWD dataset.

It can be seen from Figure 4 that HR-YOLO can still accurately detect wearing compliance for situations such as small targets and complex industrial backgrounds in the safety helmet dataset from an oil-and-gas station.

Figure 4. Some examples of detection results on an oil-and-gas station dataset.

To further verify the robustness of HR-YOLO, we compare it with different methods, namely SSD, Faster RCNN, YOLOv3, YOLOv5s, YOLOv7 and YOLOv8s, on the CUMT-HelmeT dataset. Table 3 lists the values of precision, recall, F1, mAP, Params and GFLOPs for different methods. From this table, it can be seen that our method shows the highest precision, recall, F1, mAP@0.5 and mAP@0.5:0.95 compared to the other six methods. In addition, the increases in the values of precision, recall, F1, mAP@0.5 and map@0.5:0.95 reach 3.8%, 6.1%, 5.1%, 5.6% and 7.7%, respectively, compared to the results of YOLOv5. The comparative object detection results are plotted in Figure 5, and the selected original images have challenges in the detection of small objects and multiple targets, as well as under poor lighting conditions. As can be seen from the figure, our method achieves better results than the compared methods in safety helmet detection.

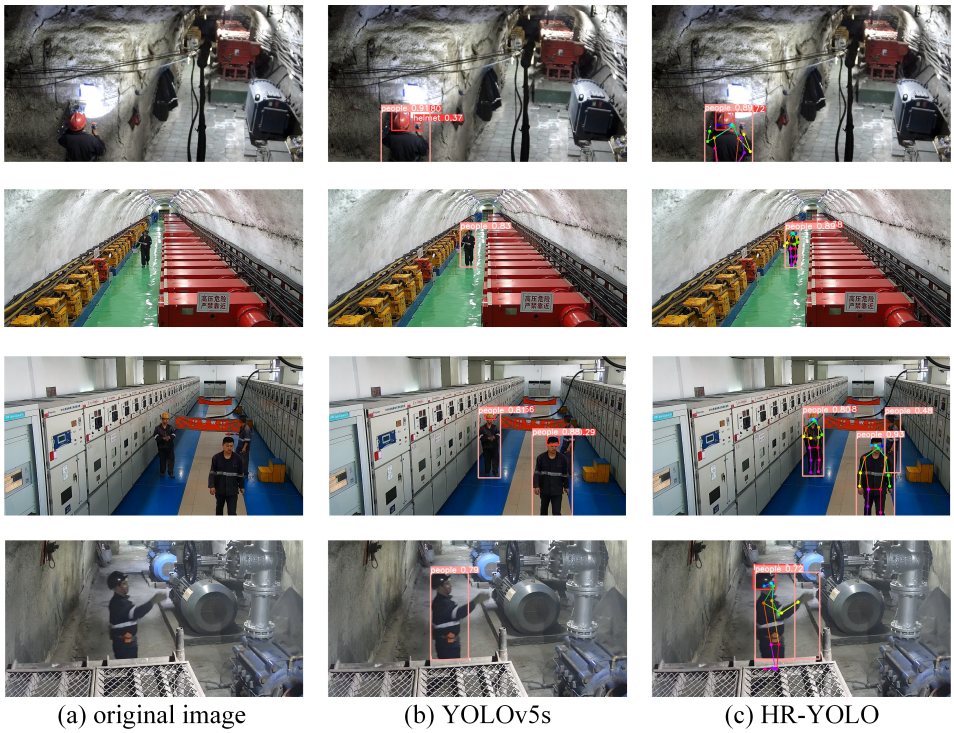


Figure 5. Some examples of detection results on the CUMT-HelmeT dataset.

Table 3. Experimental results comparing different algorithmic models on the CUMT-HelmeT dataset.

Method	P	R	F1	mAP@0.5	mAP@0.5:0.95	Params	GFLOPs
SSD	81.8	72.1	76.6	78.2	43.4	92.12 M	3.3
Faster RCNN	87.9	77.5	82.3	79.1	47.3	35.3 M	3.9
YOLOv3	86.4	83.7	85.0	88.2	60.7	61.5 M	193.8
YOLOv5s	89.6	78.6	83.7	82.3	56.9	7.3 M	15.8
YOLOv7	91.3	80.9	85.7	86.3	58.1	36.5 M	103.5
YOLOv8s	91.5	80.6	85.7	89	61.3	3.2 M	12.1
Ours	93.4	84.7	88.8	91.9	65.8	13.1 M	8.2

4.4. Model Analysis and Ablation Study

In the evaluation of LAAM on the GDUT-HWD dataset, Table 4 provides an experimental comparison between Drakent-53 as the backbone with HRNet-W32, along with the decoder of YOLOv5s and YOLOv5s + LAAM. The results demonstrate that the HR-YOLO network, specifically HRNet-W32 + YOLOv5s + LAAM, exhibits superior feature representation capabilities compared with other configurations.

Table 4. Comparison of LAAM ablation experiment results.

Backbone	Decoder	P	R	F1	mAP@0.5	mAP@0.5:0.95
Darknet-53	YOLOv5s	88.3	88.1	88.2	91.5	57.9
HRNet-W32	YOLOv5s	90.7	91.3	91.2	92.8	60.1
HRNet-W32	YOLOv5s + LAAM	92.5	91.7	91.9	93.6	61.5

Furthermore, we separately integrated OP-SGD into HRNet-W32 and the detector of YOLOv5s for comparative verification. This enables us to assess the feature aggregation capability of attention at different locations within the network, as presented in Table 5. The experimental results indicate that OP-SGD is both feasible and universally applicable in both the backbone and decoder positions.

Table 5. Comparison of OP-SGD ablation experiment results.

Backbone	Decoder	P	R	F1	mAP@0.5	mAP@0.5:0.95
HRNet-W32	YOLOv5s	90.7	91.3	91.2	92.8	60.1
HRNet-W32 +OP-SGD	YOLOv5s	90.1	93.0	91.5	94.9	61.7
HRNet-W32	YOLOv5s+OP-SGD	91.1	92.3	91.7	94.1	62.3

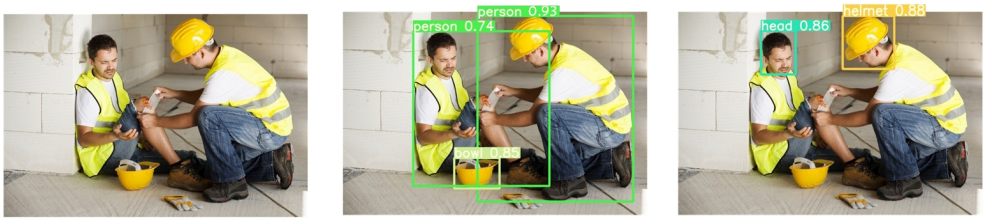
To verify the effectiveness of PA-NMS voting, we designed an ablation experiment, with the IOU and category threshold of the detection box and GT set to 0.5. As shown in Table 6, when PA-NMS voting provides key point information as guidance, the proposed algorithm improves several detection indicators in complex scenes, which indicates that the PA-NMS voting algorithm can improve detection accuracy.

Table 6. Comparison of PA-NMS ablation experiment results.

Algorithms	P	R	F1	AP@0.5					mAP@0.5	mAP@0.5:0.95
				While	Blue	Red	Yellow	None		
NMS	87.6	85.3	86.4	92.7	85.7	88.1	95.3	90.4	89.5	60.2
PA-NMS Voting	90.7	86.2	88.4	93.0	88.6	91.2	95.5	85.3	90.4	62.1

In order to obtain further evidence of the effectiveness of PDB, we conducted a comparison between HR-YOLO without PDB and HR-YOLO. As shown in Figure 6, HR-YOLO without PBD incorrectly recognizes an unworn helmet and fails to determine whether a

worker is wearing a helmet correctly, which indicates that the human posture information from PDB is important for helmet detection. In contrast, HR-YOLO showcases superior performance in this particular scenario.



(a) (b) (c)

Figure 6. Comparison of detection results between HR-YOLO without PDB and HR-YOLO. (**a**) The input from the SHWD; (**b**) the result of HR-YOLO without PDB; (**c**) the result of HR-YOLO.

4.5. Application Scenarios

To verify the safety helmet detection method proposed in this study, we apply the method to a quadrupedal robot for two real-life scenarios, namely an indoor scenario and an outdoor scenario at an oil-and-gas station. The quadruped robot system consists of a multi-depth camera, LIDAR, front view camera, router, display, power supply, built-in host and other components, as shown in Figure 7. HR-YOLO is employed in the TX2 embedded platform on quadruped robots and accelerated using TensorRt. While the robot collects environmental data, the inference speed is 32.18 frames per second, which can basically meet the requirements of on-site real-time detection. A total of 380 oil-and-gas station production scene images selected from the testing dataset were used in the safety helmet detection test.

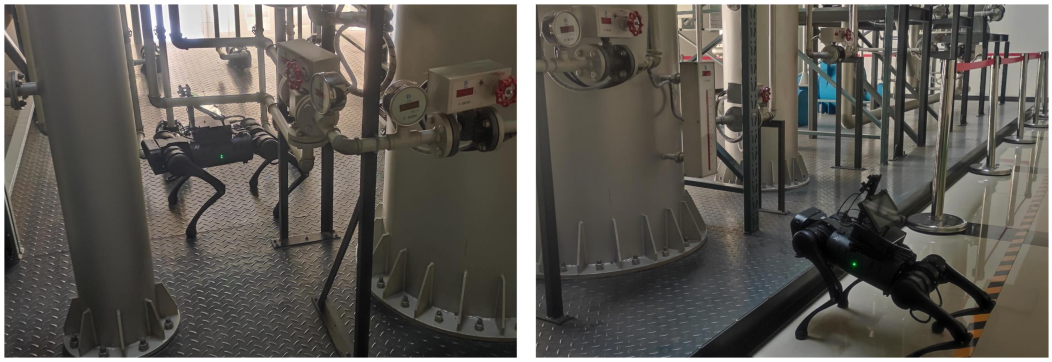


Figure 7. Schematic diagram of quadruped robot.

Table 7 compares the classification results of helmet detection in the quadruped robot system obtained with different methods. The folder size refers to the total size of all configuration files, which enables the quadruped robot system to run detection programs independently. The test accuracy and FPS of the HR-YOLO are the highest, while its folder size is still acceptable.

Table 7. Comparison of network performance in quadruped robot system.

Method	Test Accuracy	FPS	Folder Size
Yolov3	85.5	16.32	120.5
Yolov5s	86.4	23.88	14.4
Yolov7	89.3	18.65	74.8
Ours	91.5	32.18	26.3

Figure 8 shows the application of HR-YOLO in the indoor scenario. In oil-and-gas stations, workers often work near dense pipelines, and the shape of some devices is similar

to that of a safety helmet. The results demonstrate that our proposed method effectively detects instances of safety helmet wearing in such complex indoor scenes.

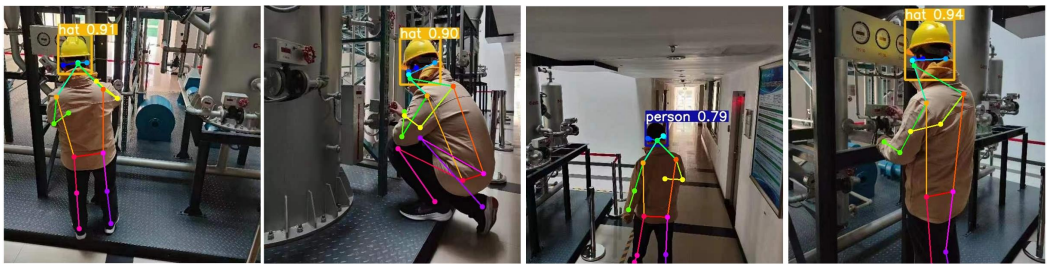


Figure 8. Application result of HR-YOLO for indoor scene detection. The colored line segments in the figure are the human pose estimation lines corresponding to the 18 key points.

Figure 9 shows the application of HR-YOLO in the outdoor scenario. It can be seen that our method can accurately detect whether workers are wearing safety helmets. The quadrupedal robot can receive detection information from HR-YOLO and transfer the information to the workers who are not wearing helmets correctly.

Figure 9. Application result of HR-YOLO for outdoor scene detection. The colored line segments in the figure are the human pose estimation lines corresponding to the 18 key points.

5. Conclusions

In this paper, we have presented a new network named HR-YOLO with a PA-NMS voting process to synthesize safety helmet wearing information from the features of helmet objects and human pose in order to address the demands of helmet detection tasks in practical applications. To overcome the decrease in detection accuracy caused by the small size of targets, we designed OP-SGD to improve the expressive ability of the network. We also propose LAAM,m which can make the YOLOv5 decoder pay more attention to the feature information from human pose to enhance network representation and suppress interference from irrelevant features. In addition, we propose a new post-processing algorithm named PA-NMS voting, which uses a suppression algorithm based on pose information constraints to determine the confidence of bounding boxes and utilizes the voting operation to obtain a new optimal bounding box. Finally, HR-YOLO was compared with other mainstream object detection method, and an ablation study has was designed to evaluate the performance of the proposed method. The experimental results indicate that HR-YOLO surpasses other algorithms in safety helmet wearing detection tasks, with commendable robustness when faced with diverse noise conditions, lighting variations and degrees of occlusion. The results also show the practical value of the proposed method in various applications. In the future, we will focus on how to explore further optimization of the network structure, incorporate multi-task output branches and enhance the network's capability to detect diverse multi-modal information. Moreover, we will further reduce compute/memory cost, improve training instability and support efficient distributed training to face scalability issues in large-scale applications.

Author Contributions: Methodology, Y.L.; Validation, J.L.; Data curation, S.D.; Writing—original draft, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grant number NSFC 61972353, NSF IIS-1816511, OAC-1910469 and Strategic Cooperation Technology Projects of CNPC and CUPB: ZLZX2020-05.

Institutional Review Board Statement: This research is licensed to allow unrestricted reuse.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are contained within the article. The GDUT-HWD dataset can be download in https://github.com/wujixiu/helmet-detection?tab=readme-ov-filel. The SHWD dataset can be download in https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset. The CUMT-HelmeT dataset can be download in https://github.com/CUMT-AIPR-Lab/CUMT-AIPR-Lab. The oil-and-gas station dataset dataset can be download in https://github.com/a23456r/IndustryHelmetDetectionDatabase.

Conflicts of Interest: Author Xingtao Li was employed by the company China National Oil and Gas Exploration and Development Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. 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, Canada, 18–22 June 2023; pp. 5349–5357.
2. Tran, D.N.N.; Pham, L.H.; Jeon, H.J.; Nguyen, H.H.; Jeon, H.M.; Tran, T.H.P.; Jeon, J.W. Robust automatic motorcycle helmet violation detection for an intelligent transportation system. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 18–22 June 2023; pp. 5340–5348.
3. Jia, W.; Xu, S.; Liang, Z.; Zhao, Y.; Min, H.; Li, S.; Yu, Y. Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector. *IET Image Process.* **2021**, *15*, 3623–3637. [CrossRef]
4. Agrahari, A.; Singh, D. Smart city transportation technologies: automatic no-helmet penalizing system. In *Blockchain Technology for Smart Cities*; Springer: Singapore , 2020; pp. 115–132.
5. Herrmann, A.; Liu, M.; Pilla, F.; Shorten, R. A new take on protecting cyclists in smart cities. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 3992–3999. [CrossRef]
6. Rubaiyat, A.H.; 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; IEEE: New York, NY, USA, 2016; pp. 135–142.
7. Tai, W.; Wang, Z.; Li, W.; Cheng, J.; Hong, X. DAAM-YOLOV5: A Helmet Detection Algorithm Combined with Dynamic Anchor Box and Attention Mechanism. *Electronics* **2023**, *12*, 2094. [CrossRef]
8. Zhang, Y.; Qiu, Y.; Bai, H. FEFD-YOLOV5: A Helmet Detection Algorithm Combined with Feature Enhancement and Feature Denoising. *Electronics* **2023**, *12*, 2902. [CrossRef]
9. Cheng, R.; He, X.; Zheng, Z.; Wang, Z. Multi-scale safety helmet detection based on SAS-YOLOv3-tiny. *Appl. Sci.* **2021**, *11*, 3652. [CrossRef]
10. Song, H. Multi-scale safety helmet detection based on RSSE-YOLOv3. *Sensors* **2022**, *22*, 6061. [CrossRef]
11. Chiverton, J. Helmet presence classification with motorcycle detection and tracking. *IET Intell. Transp. Syst.* **2012**, *6*, 259–269. [CrossRef]
12. Li, Y.; Wei, H.; Han, Z.; Huang, J.; Wang, W. Deep learning-based safety helmet detection in engineering management based on convolutional neural networks. *Adv. Civ. Eng.* **2020**, *2020*, 1–10. [CrossRef]
13. Silva, R.; Aires, K.; Santos, T.; Abdala, K.; Veras, R.; Soares, A. Automatic detection of motorcyclists without helmet. In Proceedings of the 2013 XXXIX Latin American Computing Conference (CLEI), Caracas, Venezuela, 7–11 October 2013; IEEE: New York, NY, USA, 2013; pp. 1–7.
14. E Silva, R.R.V.; Aires, K.R.T.; Veras, R.d.M.S. Helmet detection on motorcyclists using image descriptors and classifiers. In Proceedings of the 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images, Rio de Janeiro, Brazil, 26–30 August 2014; IEEE: New York, NY, USA, 2014; pp. 141–148.
15. Waranusast, R.; Bundon, N.; Timtong, V.; Tangnoi, C.; Pattanathaburt, P. Machine vision techniques for motorcycle safety helmet detection. In Proceedings of the 2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013), Wellington, New Zealand, 27–29 November 2013; IEEE: New York, NY, USA, 2013; pp. 35–40.
16. 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]
17. 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; IEEE: New York, NY, USA, 2020; pp. 1–7.

18. Guo, S.; Li, D.; Wang, Z.; Zhou, X. Safety helmet detection method based on faster r-cnn. In Proceedings of the Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, 17–20 July 2020; Proceedings, Part II 6; Springer: Berlin/Heidelberg, Germany, 2020; pp. 423–434.
19. Raj, A.V.; Manohar, N.; Dhyanjith, G. Helmet Detection using Single Shot Detector (SSD). In Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; IEEE: New York, NY, USA, 2021; pp. 1241–1244.
20. Dai, B.; Nie, Y.; Cui, W.; Liu, R.; Zheng, Z. Real-time safety helmet detection system based on improved SSD. In Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture, Manchester, UK, 15–17 October 2020; pp. 95–99.
21. Wang, W.; Gao, S.; Song, R.; Wang, Z. A safety helmet detection method based on the combination of ssd and hsv color space. In Proceedings of the IT Convergence and Security: Proceedings of ICITCS 2020; Springer: Berlin/Heidelberg, Germany, 2021; pp. 123–129.
22. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
23. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
25. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
26. 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: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
27. Jeong, J.; Park, H.; Kwak, N. Enhancement of SSD by concatenating feature maps for object detection. *arXiv* **2017**, arXiv:1705.09587.
28. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
29. Li, Z.; Zhou, F. FSSD: feature fusion single shot multibox detector. *arXiv* **2017**, arXiv:1712.00960.
30. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
31. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
32. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
33. 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.
34. 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, BC, Canada, 18–22 June 2023; pp. 7464–7475.
35. Wen, C.Y.; Chiu, S.H.; Liaw, J.J.; Lu, C.P. The safety helmet detection for ATM's surveillance system via the modified Hough transform. In Proceedings of the IEEE 37th Annual 2003 International Carnahan Conference on Security Technology, Taipei, Taiwan, 14–16 October 2003; Proceedings; IEEE: New York, NY, USA, 2003; pp. 364–369.
36. Wu, H.; Zhao, J. An intelligent vision-based approach for helmet identification for work safety. *Comput. Ind.* **2018**, *100*, 267–277. [CrossRef]
37. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
38. 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; IEEE: New York, NY, USA, 2019; pp. 2495–2499.
39. Jamtsho, Y.; Riyamongkol, P.; Waranusast, R. Real-time license plate detection for non-helmeted motorcyclist using YOLO. *ICT Express* **2021**, *7*, 104–109. [CrossRef]
40. Wu, F.; Jin, G.; Gao, M.; Zhiwei, H.; Yang, Y. Helmet detection based on improved YOLO V3 deep model. In Proceedings of the 2019 IEEE 16th International Conference on Networking, sensing and CONTROL (ICNSC), Banff, AB, Canada, 9–11 March 2019; IEEE: New York, NY, USA, 2019; pp. 363–368.
41. Zhou, F.; Zhao, H.; Nie, Z. Safety helmet detection based on YOLOv5. In Proceedings of the 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China, 22–24 January 2024; IEEE: New York, NY, USA, 2021; pp. 6–11.
42. Wang, J.; Zhou, H.; Sun, H.; Su, Z.; Li, X. A Violation Behaviors Detection Method for Substation Operators based on YOLOv5 And Pose Estimation. In Proceedings of the 2022 IEEE 3rd China International Youth Conference on Electrical Engineering (CIYCEE), Wuhan, China, 3–5 November 2022; IEEE: New York, NY, USA, 2022; pp. 1–5.

43. Tan, S.; Lu, G.; Jiang, Z.; Huang, L. Improved YOLOv5 network model and application in safety helmet detection. In Proceedings of the 2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR), Tokoname, Japan, 4–6 March 2021; IEEE: New York, NY, USA, 2021; pp. 330–333.
44. Cao, J.; Lin, Z.; Huang, G.B. Self-adaptive evolutionary extreme learning machine. *Neural Process. Lett.* **2012**, *36*, 285–305. [CrossRef]
45. Matias, T.; Araújo, R.; Antunes, C.H.; Gabriel, D. Genetically optimized extreme learning machine. In Proceedings of the 2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA), Cagliari, Italy, 10–13 September 2013; IEEE: New York, NY, USA, 2013; pp. 1–8.
46. Ma, L.; Khorasani, K. A new strategy for adaptively constructing multilayer feedforward neural networks. *Neurocomputing* **2003**, *51*, 361–385. [CrossRef]
47. Cortes, C.; Gonzalvo, X.; Kuznetsov, V.; Mohri, M.; Yang, S. Adanet: Adaptive structural learning of artificial neural networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 874–883.
48. Li, H.; Yang, Y.; Chen, D.; Lin, Z. Optimization algorithm inspired deep neural network structure design. In Proceedings of the Asian Conference on Machine Learning, Beijing, China, 14–16 November 2018; pp. 614–629.
49. Lu, Y.; Zhong, A.; Li, Q.; Dong, B. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 3276–3285.
50. Park, H.J.; Kang, J.W.; Kim, B.G. ssFPN: Scale Sequence (S 2) Feature-Based Feature Pyramid Network for Object Detection. *Sensors* **2023**, *23*, 4432. [CrossRef]
51. Zhou, B.; Liu, J.; Sun, W.; Chen, R.; Tomlin, C.J.; Yuan, Y. pbSGD: Powered Stochastic Gradient Descent Methods for Accelerated Non-Convex Optimization. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), Yokohama, Japan, 11–17 July 2020; pp. 3258–3266.
52. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
53. Ying, X.; Wang, Y.; Wang, L.; Sheng, W.; An, W.; Guo, Y. A stereo attention module for stereo image super-resolution. *IEEE Signal Process. Lett.* **2020**, *27*, 496–500. [CrossRef]
54. 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.
55. 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, 18–22 June 2018; pp. 7132–7141.
56. Anwar, S.; Barnes, N. Densely residual laplacian super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1192–1204. [CrossRef]
57. Han, G.; Zhu, M.; Zhao, X.; Gao, H. Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection. *Comput. Electr. Eng.* **2021**, *95*, 107458. [CrossRef]
58. Chen, S.; Lan, J.; Liu, H.; Chen, C.; Wang, X. Helmet wearing detection of motorcycle drivers using deep learning network with residual transformer-spatial attention. *Drones* **2022**, *6*, 415. [CrossRef]
59. Cantarini, G.; Tomenotti, F.F.; Noceti, N.; Odone, F. HHP-Net: A light Heteroscedastic neural network for Head Pose estimation with uncertainty. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 3521–3530.
60. Gidaris, S.; Komodakis, N. Object detection via a multi-region and semantic segmentation-aware cnn model. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1134–1142.

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.