electronics

MDPI

Article

A Safety Helmet Detection Model Based on YOLOv8-ADSC in Complex Working Environments

Jingyang Wang [1,†], Bokai Sang [1], Bo Zhang [2,*,†] and Wei Liu [1]

[1] Hebei University of Science and Technology, Shijiazhuang 050018, China; jingyangw@hebust.edu.cn (J.W.); kk_542132700@163.com (B.S.); lw17732712884@163.com (W.L.)
[2] Hebei University of Engineering Science, Shijiazhuang 050091, China
[*] Correspondence: yatouzhangbo@163.com
[†] These authors contributed equally to this work.

Abstract: A safety helmet is indispensable personal protective equipment in high-risk working environments. Factors such as dense personnel, varying lighting conditions, occlusions, and different head postures can reduce the precision of traditional methods for detecting safety helmets. This paper proposes an improved YOLOv8n safety helmet detection model, YOLOv8-ADSC, to enhance the performance of safety helmet detection in complex working environments. In this model, firstly, Adaptive Spatial Feature Fusion (ASFF) and Deformable Convolutional Network version 2 (DCNv2) are used to enhance the detection head, enabling the network to more effectively capture multi-scale information of the target; secondly, a new detection layer for small targets is incorporated to enhance sensitivity to smaller targets; and finally, the Upsample module is replaced with the lightweight up-sampling module Content-Aware ReAssembly of Features (CARAFE), which increases the perception range, reduces information loss caused by up-sampling, and improves the precision and robustness of target detection. The experimental results on the public Safety-Helmet-Wearing-Dataset (SHWD) demonstrate that, in comparison to the original YOLOv8n model, the mAP@0.5 of YOLOv8-ADSC has increased by 2% for all classes, reaching 94.2%, and the mAP@0.5:0.95 has increased by 2.3%, reaching 62.4%. YOLOv8-ADSC can be better suited to safety helmet detection in complex working environments.

Keywords: YOLOv8n; safety helmet detection; feature fusion; deformable convolutional networks; new detection layer; up-sampling

check for updates

Citation: Wang, J.; Sang, B.; Zhang, B.; Liu, W. A Safety Helmet Detection Model Based on YOLOv8-ADSC in Complex Working Environments. *Electronics* **2024**, *13*, 4589. https://doi.org/10.3390/electronics13234589

Academic Editor: Prashan Premaratne

Received: 10 October 2024
Revised: 7 November 2024
Accepted: 18 November 2024
Published: 21 November 2024

1. Introduction

In high-risk working environments such as construction, mining, and industrial production, safety helmets are vital as the most basic and crucial form of protective tool, significantly reducing the likelihood of head injuries. Statistics indicate that a substantial percentage of workplace accidents occur due to head injuries, emphasizing the necessity for effective protective measures. Due to the influence of complex and changeable working environments, the precision of traditional methods for detecting safety helmets decreases, and enhancing the detection performance of safety helmets in complex working environments is urgently needed. These complex working environments involve dense personnel, changes in lighting conditions, occlusion problems, head posture changes, diverse background environments, operation dynamics, and diversity of protective equipment. Therefore, the detection methods need higher robustness and adaptability and can accurately identify and locate safety helmets under unfavorable conditions, thereby contributing to creating a safer working environment and ultimately reducing the incidence of work-related injuries.

In the early days, traditional manual feature extraction was used to detect safety helmets. Silva et al. [1] used Circle Hough Transform (CHT) and Histogram of Oriented Gradient (HOG) descriptors to locate pedestrians and safety helmets. Then, they used

Multi-Layer Perceptron (MLP) classifiers to detect safety helmets. The main implementation steps of this traditional algorithm include image preprocessing, region detection of safety helmets, feature extraction, and target classification. Due to its simple structure, the traditional algorithm is highly computationally efficient under limited computing resources [2]. However, its detection accuracy is low, and it performs poorly when processing images with complex backgrounds and light variations [3].

As deep learning has undergone significant advancements and rise, particularly the remarkable achievements of Convolutional Neural Network (CNN) [4], the selection of algorithms for detecting safety helmets has become richer and more diverse.

Deep learning target detection algorithms can be categorized into two distinct classes: two-stage approaches and one-stage approaches [5]. Region-CNN (R-CNN) [6] represents a seminal two-stage target detection algorithm. It is structured into two main stages. In the first stage, a group of candidate regions that may contain targets are extracted in the image using region extraction technology. In the second stage, feature extraction and classification recognition are carried out on these regions, and regression algorithms are used to locate the target's location accurately. Subsequently, two-stage target detection algorithms, Fast-R-CNN [7] and Faster-R-CNN [8], were proposed successively. Although they can achieve higher accuracy, their detection speed is slow, and they are difficult to deploy in working environments with limited computing resources [9]. One-stage typical target detection algorithms include RetinaNet [10] and You Only Look Once (YOLO) series algorithms [11–14]. Its main feature is that it directly performs bounding box regression and target classification and obtains the target detection results at one time. This end-to-end design makes the one-stage algorithm more concise and efficient when dealing with target detection tasks. Two-stage methods usually require the generation of candidate boxes first, followed by classification and position regression of the candidate boxes, so two independent stages are required to complete the target detection, which increases the computational complexity and processing time. Therefore, the one-stage algorithm has greater advantages in working environments with high real-time requirements.

According to the requirement of accuracy and real-time detection of safety helmets in working environments characterized by constrained computing resources, we propose the YOLOv8-ADSC model, an improved version of YOLOv8n. This model endeavors to address the limitations of traditional target detection models when applied to the complex environments of construction sites, thereby enhancing both the efficiency and accuracy of safety helmet detection.

As deep learning technology continues to advance, both the one-stage and the two-stage algorithms have achieved remarkable results in the research of safety helmet detection. Wang [15] used ResNet as the feature extraction network for Faster R-CNN; optimized the generation of candidate regions to deal with the challenges of the current deep learning algorithms, which struggle due to their inadequate speed and accuracy when it comes to detecting safety helmets; and built a helmet-wearing detection system to alert individuals neglecting to wear safety helmets. However, the proposed method still faces challenges in real-time detection due to its computational cost and slower inference speed in complex environments. Huang et al. [16] proposed an enhanced safety helmet detection algorithm, building upon the YOLOv3 algorithm. This algorithm effectively enhanced the detection accuracy and speed by increasing the scale of the feature map, improving the prior frame, and using the Complete Intersection-Over-Union (CIoU) loss function. Xu et al. [17] proposed a Dual-Key Transformer Network (DKTNet) to detect small targets in complex scenarios. This network is based on a channel-wise self-attention mechanism, in which a dual-key strategy is used to enhance the correlation between the query (Q) and value (V), and the convolutional computation replaces the fully-connected computation of Q, key (K), and V in the transformer architecture. This work plays an important role in the detection of small targets, particularly in helmet detection. Wang et al. [18] proposed a safety helmet-wearing detection model based on YOLO-M. The model used the MobileNetv3 backbone network, altered the feature fusion process, and introduced a new two-way attention module at the neck of the model to enable it to obtain effective features more accurately. The experimental results on the SHWD public dataset showed that the mean Average Precision (mAP) was improved. However, there are still cases of missed detections in complex construction environments. Zhao et al. [19] proposed a safety helmet detection algorithm based on BDC-YOLOv5, which added additional detection layers and a Convolutional Block Attention Module (CBAM) to the neck of the model. Finally, a Bidirectional Feature Pyramid Network (BiFPN) was introduced into the structure of neck feature fusion, which further improved the model's accuracy and applicability, especially enhancing the detection capability of minute targets within complex environments. However, the complexity of the network may hinder its real-time application in certain scenarios. The YOLOv8n-SLIM-CA algorithm proposed by Lin added the Coordinate Attention (CA) mechanism to the backbone network and adopted the Slim-Neck structure to fuse different-sized features in the neck network [20]. The model enhanced the detection accuracy in the safety helmet detection of small targets at long distances. Song and Wang proposed a Recurrent Bidirectional Feature Pyramid Detector (RBFPDet) for detecting safety helmets [21]. The model first processed the detected images through the EfficientNet-B0 backbone network, proposed a Recurrent Bidirectional Feature Pyramid Network (RBFPN) for feature fusion, and finally used non-local feature enhancement blocks for non-local feature redistribution. Gao et al. focused on fault diagnosis of rotating machinery under different operating conditions, presenting the Domain Feature Decoupling Network (DFDN) [22] and the Multi-source Domain Information Fusion Network (MDIFN) [23]. DFDN's core feature is decoupling extracted features into domain-related and fault-related components, which are then adaptively integrated using a significance fusion method. In contrast, the MDIFN leverages adversarial transfer learning and fine-grained information fusion to effectively combine feature information from multiple source domains. Both methods showcase innovative approaches and convincing experimental results, advancing research and applications in the field of object detection.

In summary, the YOLO algorithm of the one-stage model had better detection accuracy and a faster detection speed, which made it stand out among the algorithms for detecting safety helmet wearing. However, the detection effect still had certain limitations in gloomy light, at different distances, and in complex backgrounds and environments [24]. Therefore, this paper took the safety helmet detection problem in complex working environments as the key research focus.

2. Methods

2.1. YOLOv8n

YOLOv8 is an algorithmic model capable of performing image classification, target detection, and instance segmentation. Building on the success of previous YOLO versions, it incorporates new functions and improvements. YOLOv8 represents one of the most effective models in the YOLO series, available in five versions (n, s, m, l, and x), ranging in model size from small to large. These versions vary in depth and width, allowing the model to adapt to diverse application scenarios, from lightweight to high-precision tasks. In this paper, we choose the lightweight YOLOv8n as our base model, considering model size and computational efficiency. YOLOv8n is well suited for embedded systems and mobile devices, making it more practical for deployment in environments with limited computing resources. The original network architecture of YOLOv8n is illustrated in Figure 1; the input part receives the original image data and passes it to the backbone for processing. The input image will undergo preprocessing and enhancement operations to meet the model's requirements. The backbone uses the Darknet-53 network structure for feature extraction, enabling it to capture essential information from the image that subsequent network layers can then utilize. It follows the Spatial Pyramid Pooling Fast (SPPF) structure in YOLOv5. The neck part uses the Path Aggregation Network–Feature Pyramid Network (PAN-FPN) for feature fusion [25]. This fusion methodology augments the target detection model's capability to perceive targets, as well as the details and contextual information associated with those targets. The head part adopts a modern decoupling head architecture, allowing for the segregation of classification and detection heads. However, there are still many problems in detecting safety helmets, such as dense personnel, mutual occlusion, and multi-scale detection of targets under gloomy light. To address the aforementioned troubles, this paper focuses on a series of improvements based on YOLOv8n to further improve its detection performance and model generalization ability.

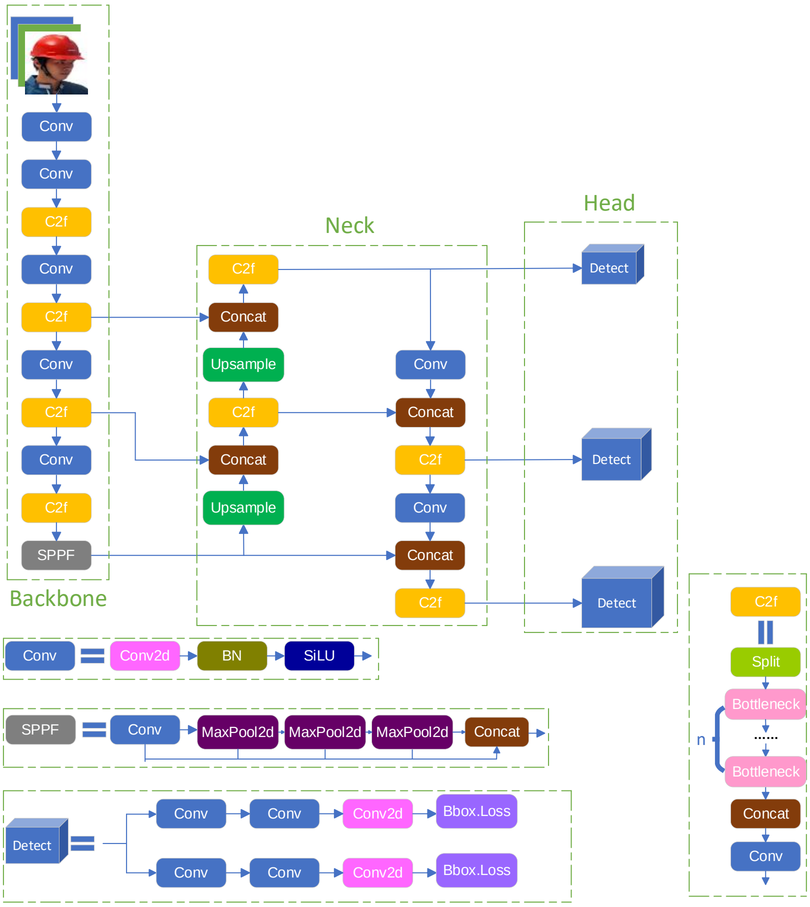


Figure 1. Framework of YOLOv8n original network. Different colors represent different modules.

2.2. YOLOv8-ADSC

The network architecture of YOLOv8-ADSC, a model that builds upon the foundation of YOLOv8n, is shown in Figure 2. The parts highlighted within the red solid lines in Figure 2 represent the innovative improvements in the model. Firstly, to improve the model's capacity to discern targets of varying scales and solve the inconsistency problem between different scales, we use ASFF [26] and DCNv2 [27] to improve the detection head. Secondly, to more effectively detect people wearing safety helmets at a distance and in dense situations, we add a new layer for detecting small targets to the network. Finally, we replace the traditional Upsample module with a lightweight up-sampling CARAFE [28] module, which can increase the perception range to improve its overall detection capability.

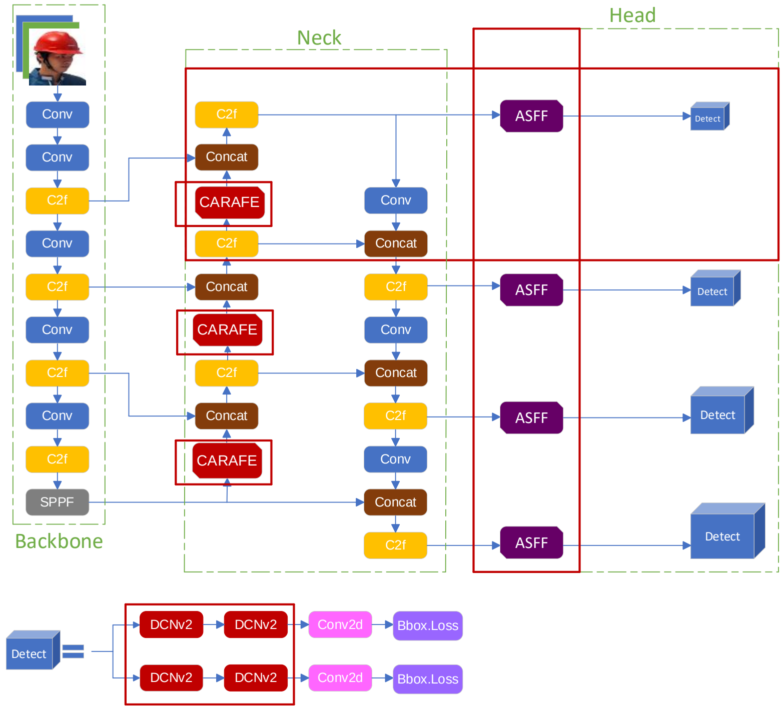


Figure 2. Framework of YOLOv8-ADSC network.

2.2.1. Introduce ASFF and DCNv2 to Improve the Detection Head

In complex working environments with background interference, character overlap, and severe gloomy light, detecting wearing safety helmets requires a better feature fusion network. The traditional target detection network usually uses only a single layer of the feature map and ignores the feature map information of other layers when the feature map matches the object. To address the issue of inconsistency between different scales, we introduced ASFF in the head of YOLOv8 to effectively filter conflicting information and improve the multi-scale fusion of image features. Consequently, the network can better capture multi-scale information of the target.

The ASFF network utilizes multi-level features extracted from the backbone network, enabling the network to better understand the image through an adaptive weighted fusion mechanism. The working principle of ASFF is illustrated in Figure 3. ASFF adaptively learns the fusion space weight of each scale feature map, which is divided into two steps: identically rescaling and adaptive fusion. Identically rescaling ensures that all feature maps have the same spatial size in preparation for subsequent adaptive fusion. The identically rescaling operations are as follows:

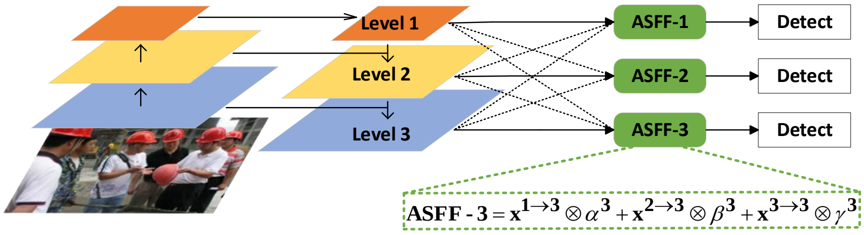

Level 1
Level 2
Level 3
ASFF-1
ASFF-2
ASFF-3
Detect
Detect
Detect
$\mathrm{ASFF-3} = x^{1\to3} \otimes \alpha^3 + x^{2\to3} \otimes \beta^3 + x^{3\to3} \otimes \gamma^3$

Figure 3. Working principle of ASFF.

For ASFF-1: We apply the 3 × 3 convolution operation to the feature map of level2, and the stride is 2, which halves the size of the feature map to obtain $x^{2\to1}$. For the feature map of level3, we apply the 3 × 3 MaxPool operation, the stride is 2, and then we apply the 3 × 3 convolution operation, the stride of which is also 2, which will reduce the size of the feature map by four times to obtain $x^{3\to1}$.

For ASFF-2: We apply the 3 × 3 convolution operation to the feature map of level3, and the stride is 2, which halves the size of the feature map to obtain $x^{3\to2}$. We apply the 1 × 1 convolution operation to the feature map of level1 and resize to twice the feature map of level1 to obtain $x^{1\to2}$.

For ASFF-3: We apply the 1 × 1 convolution operation to the feature map of level2 and resize to twice the feature map of level2 to obtain $x^{2\to3}$. We apply the 1 × 1 convolution operation to the feature map of level1 and resize to four times the feature map of level1 to obtain $x^{1\to3}$.

Those abovementioned identically rescaling operations aim to scale the feature maps of different levels for the next adaptive fusion process. The dotted line box in Figure 3 describes how to perform an adaptive fusion of features with ASFF-3, where x^1, x^2, and x^3 represent features from level1, level2, and level3. The new fusion feature ASFF-3 is obtained by multiplying with the feature weights α^3, β^3, and γ^3 from different layers and then adding. $x_{ij}^{n\to l}$ signifies the feature vector adjusted from the n-th layer to the position (i, j) of the l-th layer, $l = 1, 2, 3$. The calculation process of the corresponding feature fusion y_{ij}^l of the l-th layer is shown in Equation (1).

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{1\to l} + \beta_{ij}^l \cdot x_{ij}^{2\to l} + \gamma_{ij}^l \cdot x_{ij}^{3\to l} \tag{1}$$

$\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$, $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l \in [0, 1]$, and satisfies Equation (2), where $\lambda_{\alpha_{ij}}^l$, $\lambda_{\beta_{ij}}^l$, and $\lambda_{\gamma_{ij}}^l$ are obtained when backpropagating and updating the network and are weighted scalar maps calculated from $x^{1\to l}$, $x^{2\to l}$, and $x^{3\to l}$ using a 1 × 1 convolution layer, respectively.

$$\alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}} \tag{2}$$

In the original YOLOv8 detection head, a standard convolution module with a constant nuclear shape/size is applied, resulting in an immutable receptive field. Considering that the safety helmet of the worker to be detected in this paper will always change with the character, the posture and observation angle will change, which brings difficulties to the detection. In addition, the original YOLOv8 tends to produce missed and false detection results when there are overlaps between different targets. To address the problems above, in previous work, deformable convolution is generally directly added to the backbone network to deal with them [29]. However, we innovatively replaced the standard convolution in the detection head module with the DCNv2 module without changing the internal structure of the deformable convolution. Figure 4 illustrates the comparison between ordinary convolution and DCNv2.

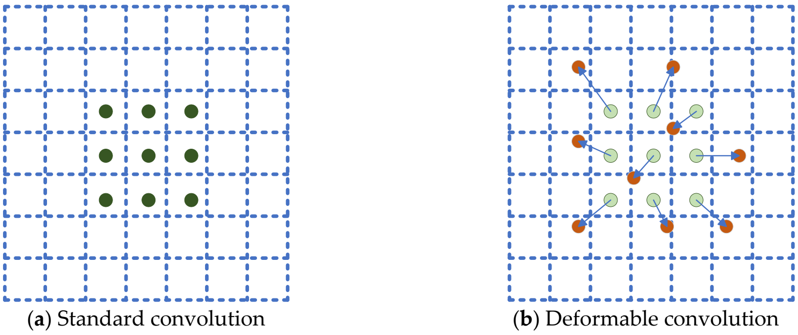

(**a**) Standard convolution (**b**) Deformable convolution

Figure 4. Comparison of standard convolution and deformable convolution DCNv2. Green circles represent sampling points for standard convolution, and orange circles represent sampling points for deformable convolution.

In this way, DCNv2 enables adaptive adjustment of its receptive field according to the contour and size of the target. The sampling can be tailored more closely to the shape and size of the target, which meets the requirement of feature extraction for moving and overlapping objects and is more robust. Figure 5 demonstrates the feature extraction process employed by the deformable convolutional module.

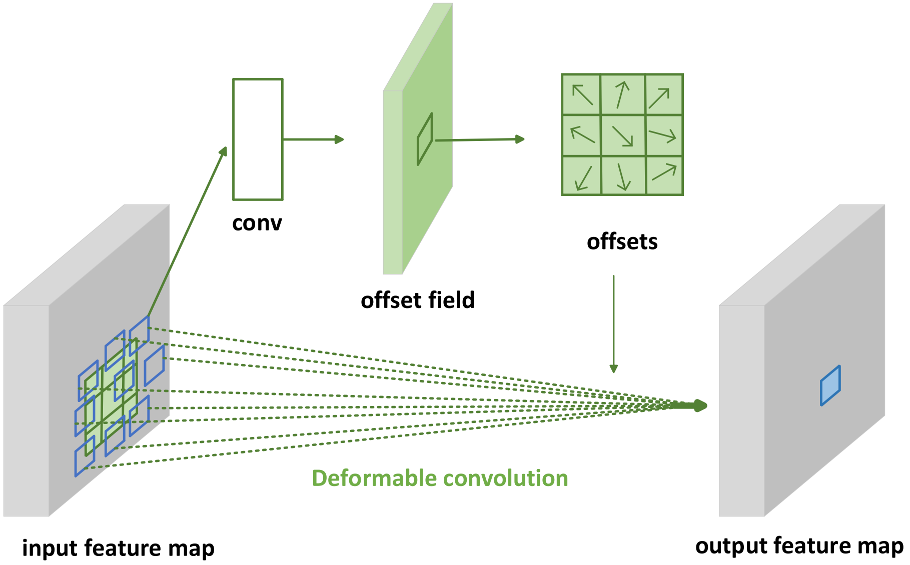


Figure 5. Feature extraction process of DCN.

Deformable convolution obtains the offset by adding a convolution layer onto the input feature layer, that is, allowing the sampling grid to deform freely, changing the receptive field by additional parameters, and learning convolution kernels and offsets used to generate output features simultaneously during training. For example, given a convolution kernel at K sampling positions, k is an integer, and $P_k \in \{(-1,-1),(-1,0),\ldots,(1,0),(1.1)\}$ defines a 3 × 3 convolution. DCNv2 is shown in Equation (3):

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \tag{3}$$

In the formula, y is the output feature map, x is the map of the input feature map, p is the real pixel coordinate in y, and p_k is the position of the convolution kernel. Using Δp_k to represent the learnable offset at the k position, Δm_k is the learned weight coefficient.

By introducing ASFF and DCNv2 into the YOLOv8 detection head, the network's capacity to perceive targets across varying scales is enhanced, and the feature information of irregular targets can be better learned and extracted, further improving the model performance.

2.2.2. Add a New Layer for Detecting Small Targets

One key factor contributing to the suboptimal performance of small-target detection in complex working environments lies in the minute size of small-target samples. Although the YOLOv8n network has three different detection heads (80 × 80, 40 × 40, and 20 × 20) to detect small, medium, and large targets, respectively, the feature map obtained by the 80 × 80 detection head is generated through multiple convolutions and down-sampling fusion of the network. In this process, a lot of shallow feature information is lost, and it is difficult to extract meaningful feature information of small targets from the generated feature maps. Thus, a 160 × 160 detection head is proposed to join the shallow and deep feature maps for detection. Specifically, an additional small-target detection layer is added to the initial three detection layers to improve the detection of small targets. This addition deepens the network, enabling feature information extraction into deeper layers.

It can also better integrate multi-scale information to understand different sizes of targets more comprehensively and enhance the detection performance across multi-scale targets. Concurrently, this approach enhances the network's comprehension of target contexts and reduces issues of false positives and missed detections. Directing the network's attention toward smaller targets improves the detection accuracy for these smaller targets. Figure 6 shows the modified model's feature fusion network architecture, incorporating a small-target detection layer.

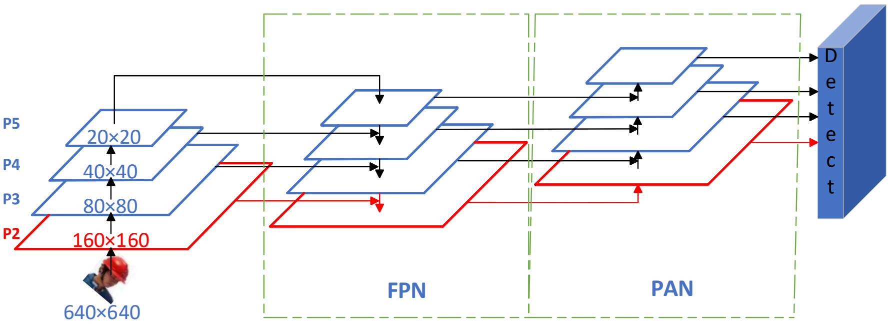


Figure 6. Feature fusion network with adding additional target detection layer.

The ASFF introduced in Section 2.2.1 is only adapted to the case of the three detection heads owned by the original model, which limits its application to more complex detection tasks. For this reason, we have expanded and innovated the ASFF module so that it can adapt to the needs of four detection heads. Following integrating a small-target detection layer, the ASFF module within the detection head manifests, as illustrated in Figure 7. This improvement retains the adaptive characteristics of ASFF and augments the network's capacity to process multi-scale features, making it perform better in more complex detection tasks.

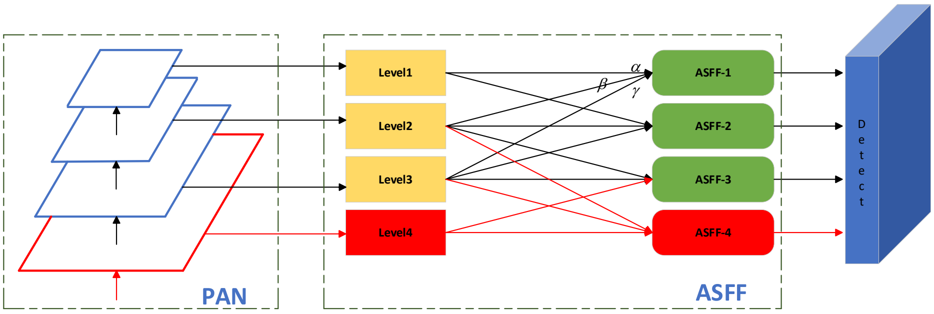


Figure 7. ASFF after incorporating a new small-target detection layer.

2.2.3. Replace the Upsample Module with the CARAFE Module

Feature up-sampling is a key operation in the YOLOv8 network. The feature map up-sampling of the original YOLOv8 network is processed by the Upsample module, which adopts the nearest neighbor interpolation up-sampling algorithm. The algorithm solely relies on pixel spatial positions to determine the up-sampling kernel, neglecting semantic information from the feature map. This results in up-sampling with a limited perceptual field. Therefore, the CARAFE module replaces the Upsample module in this paper. The working principle of CARAFE is illustrated in Figure 8.

The up-sampling process involves calculating the dot product between the up-sampling kernel at each position and the pixels within its corresponding neighborhood in the input feature map. This process is known as feature recombination. CARAFE up-sampling features a large receptive field for feature recombination, guiding the process based on input characteristics while maintaining a lightweight design. CARAFE comprises two key modules: the up-sampling kernel prediction module above Figure 8 and the feature recombination module below Figure 8. Here is how CARAFE works.

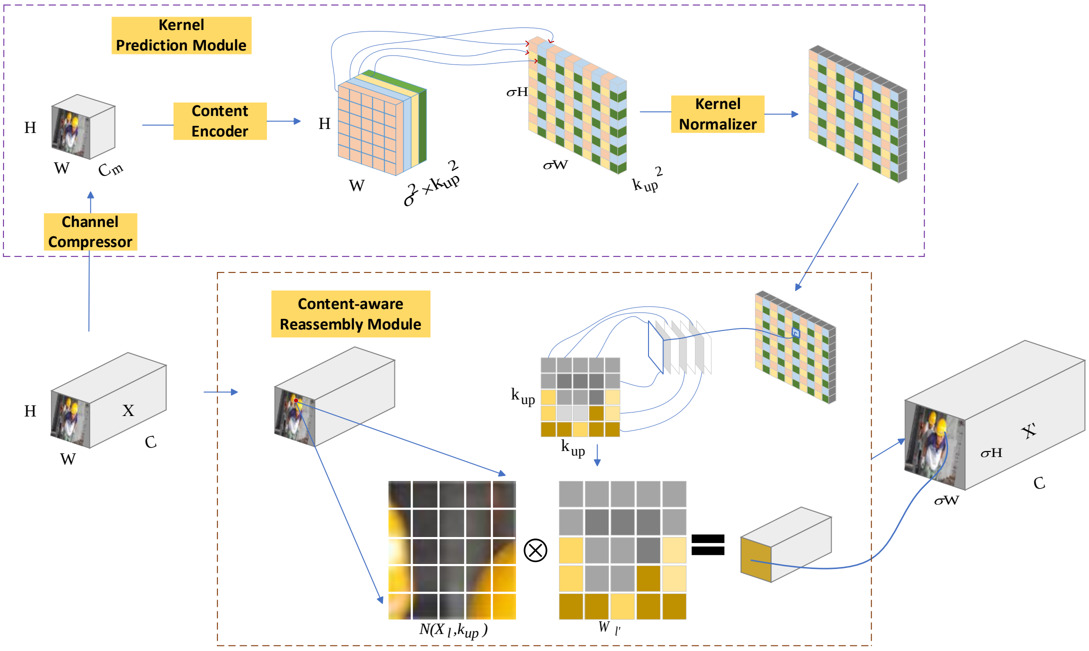


Figure 8. Working principle of CARAFE.

Assuming that the up-sampling magnification is σ and an input feature graph X with the shape of $H \times W \times C$ is given, CARAFE initially employs the up-sampling kernel prediction module to forecast the up-sampling kernel. Subsequently, it utilizes the feature recombination module to accomplish the up-sampling process, resulting in the output feature map X' with a specified shape of $\sigma H \times \sigma W \times \sigma C$. The up-sampling kernel prediction module is divided into three steps. Firstly, a 1×1 convolution is used to compress the channel number of the input feature graph from the original C to C_m, reducing the subsequent calculation amount. Secondly, a convolution layer of $k_{encoder} \times k_{encoder}$ is used to predict the up-sampling kernel for the compressed $H \times W \times C_m$ feature map, and the up-sampling kernel with the shape $\sigma H \times \sigma W \times k_{up}^2$ is generated, where C_m is the number of input channels, and $\sigma^2 k_{up}^2$ is the number of output channels. k_{up}^2 represents the size of the up-sampling kernel. Finally, softmax is used to normalize the predicted up-sampling kernel. In the following feature recombination module, the eigenvalue X_l from the input feature graph X is taken, where l is a position at the coordinates (i, j). $l' = N(X_l, k_{up})$, which is a sub-square region of $k_{up} \times k_{up}$ centered on X_l. $X'_{l'}$ is the output eigenvalue at the l' region position in X', and its calculation process is shown in Equation (4).

$$X'_{l'} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} W_{l'(n,m)} \cdot X_{(i+n,j+m)} \tag{4}$$

where $r = \lfloor k_{up}/2 \rfloor$, and $W_{l'}$ is the kernel weight obtained by the kernel prediction module. Finally, the output feature graph X' of shape $\sigma H \times \sigma W \times C$ is obtained.

A larger up-sampling kernel represents a larger receptive field and a corresponding increase in computational requirements. The experiments of each parameter quantity are performed to explore the more appropriate k_{up} and $k_{encoder}$ parameters, as shown in Table 1. Only by increasing k_{up} and $k_{encoder}$ simultaneously can the performance be improved. Finally, in this paper, we choose k_{up} = 3 and $k_{encoder}$ = 5 after considering the performance and the calculation amount.

Table 1. Comparative experiment of different parameters in the CARAFE module.

k_{up}	$k_{encoder}$	mAP@0.5 (%)	mAP@0.5:0.95 (%)	GFLOPs
1	3	92.8	60.8	8.2
3	5	93.1	61	8.4
5	7	93	60.9	9.4

3. Experiment

3.1. Experimental Environment

The experiments are conducted on a Windows 10 operating system using a 12th Gen Intel (R) Core (TM) i5-12400F six-core processor 2.50 GHz and equipped with a GPU of NVIDIA GeForce RTX 4060 Ti. Using Python 3.9 as the locale and the PyTorch deep learning framework, the epoch is set to 400.

3.2. Experimental Dataset

The SHWD [30] public dataset is utilized for model training, validation, and testing for the research experiment. This dataset, known for its use in safety helmets and human head detection, comprises 7581 images, with 9044 labeled individuals wearing helmets (positive samples) and 111,514 examples of normal heads without helmets (negative samples). The dataset includes various real-world scenarios where safety helmets are worn in intricate working environments. The dataset is divided into a training–validation set (6064 images) and a test set (1517 images), maintaining an 8:2 ratio. Furthermore, the training–validation set is further split into a training set (5457 images) and a validation set (607 images), with a 9:1 ratio.

3.3. Ablation Experiment

In this section, an ablation experiment based on the YOLOv8n model is designed to verify the effectiveness of three improvement strategies to enhance the performance of the model detection.

The abbreviations AD, S, and C, respectively, represent the improvement in the detection head mentioned in Section 2.2, the addition of a new small-target detection layer, and the replacement of the Upsample module with the CARAFE module. Table 2 shows the ablation experiment results. Figure 9 compares the accuracy and recall rates achieved by various models, while Figure 10 compares their mAP@0.5 and mAP@0.5:0.95 values.

After improving the detection head, the mAP@0.5 increased by 0.6%, and the mAP@0.5:0.95 increased by 1.1%. With the addition of a new small-target detection layer, the mAP@0.5 increased by 0.9%. After replacing the Upsample module with CARAFE, the mAP@0.5 and mAP@0.5:0.95 increased by 0.9%. The combination of the three improved methods is the best, and among the two combined improvement experiments, improving the detection head while incorporating a small-target detection layer yielded optimal results. The comparison of the PR curves of the original YOLOv8n and YOLOv8-ADSC is shown in Figure 11. The PR curves reflect the tradeoff between precision and recall. The closer the curve is to the upper right corner, the better the model's performance. We can judge by the balance point; the commonly used balance point is F1, and a higher F1 value indicates that the classification model has achieved a better balance between precision and recall. The balance point is P = R, where the slope is equal to 1. Compared with the original YOLOv8n, the YOLOv8-ADSC model improves the detection performance for wearing a safety helmet. This improvement is achieved with only a minimal increase in computational cost, which remains within an acceptable range for practical applications. Furthermore, the model sustains a frame rate of 51 Frames Per Second (FPS), thereby fully satisfying real-time processing requirements. This balance of enhanced accuracy and computational efficiency highlights the model's applicability in complex working environments.

Table 2. Results of ablation experiment.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95(%)	GFLOPs	FPS	Parameters (M)
YOLOv8n	90	87	92.2	60.1	8.2	98	3.01
YOLOv8-AD	90.7	88	92.8	61.2	8.6	78.1	4.69
YOLOv8-S	90.1	87.9	93.1	60.4	12.2	80.9	2.92
YOLOv8-C	90.9	87.8	93.1	61	8.4	70.4	3.14
YOLOv8-AD-S	91.6	88.5	93.9	62	12.9	62.9	5.07
YOLOv8-S-C	91.2	88.8	93.8	61.6	13.6	61	3.12
YOLOv8-AD-C	90.2	88.9	93.4	61.7	9.2	59	5.26
YOLOv8-ADSC	91.6	89	94.2	62.4	13.9	51	5.27

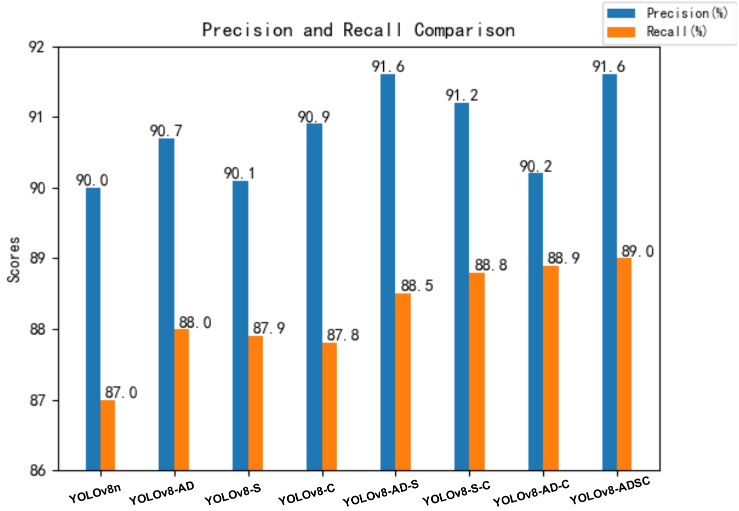


Figure 9. Comparison of the accuracy and recall rates for different models.

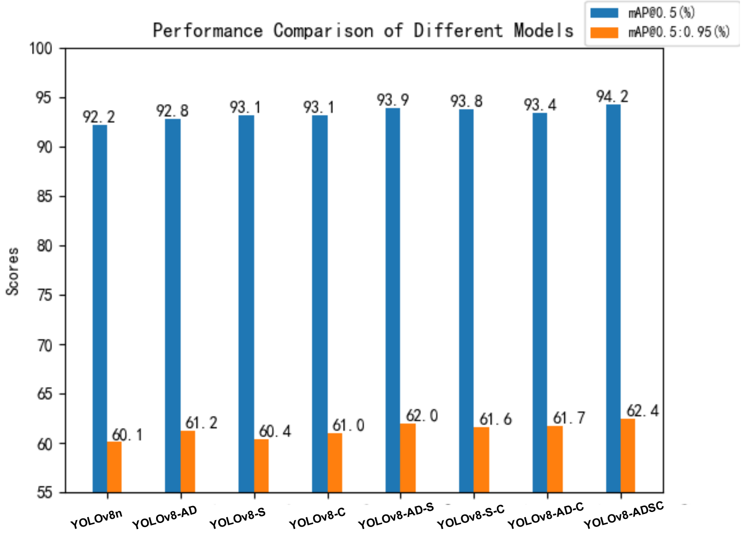
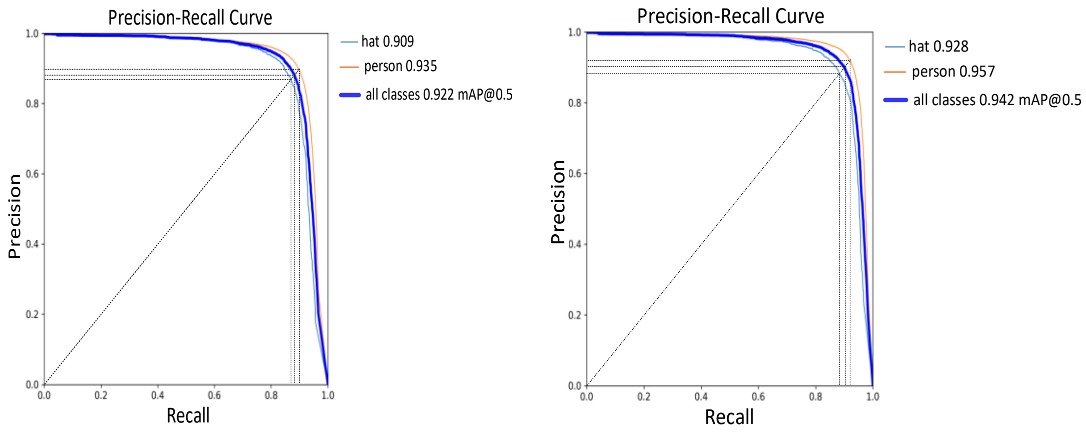


Figure 10. Comparison of mAP@0.5 and mAP@0.5:0.95 values for different models.



(**a**) YOLOv8n (**b**) YOLOv8-ADSC

Figure 11. PR curves of the original YOLOv8n and YOLOv8-ADSC. The dotted lines with a slope of 1 are used to identify the balance points of the curves.

3.4. Comparative Experiment

To validate the detection capabilities of the YOLOv8-ADSC model, we employ comparative experiments with several classic deep learning-based target detection algorithms, utilizing identical software and hardware environments and the dataset (public dataset SHWD), using precision, recall, mAP@0.5, mAP@0.5: 0.95, GFLOP, FPS (a key indicator of model real-time performance), and parameter indicators for comparison. Table 3 compares the YOLOv8-ADSC model's experimental results against classical target detection models. In contrast to the classic two-stage model Faster-R-CNN (resnet50 + FPN), YOLOv8-ADSC demonstrates notable advantages in terms of the mAP and GFLOPs. In comparison with the lightweight classical model, YOLOv8-ADSC achieves a remarkable enhancement across all performance metrics in contrast to the mainstream model of the YOLOv5 series, compared with YOLOv7-tiny, where the precision is basically the same, recall is 2.4% higher, mAP@0.5 is 1.3% higher, and mAP@0.5: 0.95 is 4.2% higher. Compared with YOLOv8s, the precision is 1.3% higher, and the model parameters are reduced by 52.7% when the mAP is slightly higher than YOLOv8s.

Table 3. Comparative experiments between YOLOv8-ADSC and traditional model.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95(%)	GFLOPs	FPS	Parameters (M)
Faster-R-CNN(ResNet50+FPN)	90.9	78.4	90.9	59.2	180	16	42
YOLOv4-tiny	85.9	76.4	81.3	46.8	6.79	60.6	5.89
YOLOv5n	90.7	86	91.3	57.8	4.1	96	1.9
YOLOv5s	90.9	87.3	92.3	59.7	15.8	90	7.03
YOLOv7-tiny	91.5	86.6	92.9	58.2	13	50.5	6.01
YOLOv8s	90.3	89.5	93.7	62.2	28.4	90.9	11.13
YOLOv8n	90	87	92.2	60.1	8.2	98	3.01
YOLOv10n	89.8	85.9	91.6	58.4	8.2	97	2.7
YOLOv8-ADSC	91.6	89	94.2	62.4	13.9	51	5.27

To expand our evaluation, we also conduct comparative experiments with the research results of other researchers on the SHWD dataset, and the experimental results are illustrated in Table 4. The experiments reveal that the YOLOv8-ADSC model surpasses other improved models, boasting a higher mAP@0.5 score, and Figure 12 shows a line chart comparing the mAP@0.5 values for all the models.

The above comparative experiments show that compared with other mainstream algorithms in complex working environments, the YOLOv8-ADSC model has better detection performance.

Table 4. Comparative experiments of YOLOv8-ADSC and other researchers' target detection models on the SHWD dataset.

Model	Image Size	mAP@0.5 (%)
RBFPDet [21]	512 × 512	89.6
DKTNet [17]	224 × 224	92.8
BDC-YOLOv5 [19]	640 × 640	93.9
YOLO-M [18]	640 × 640	93.87
YOLOv8n-SLIM-CA [20]	640 × 640	93.8
YOLOv8-ADSC	640 × 640	94.2

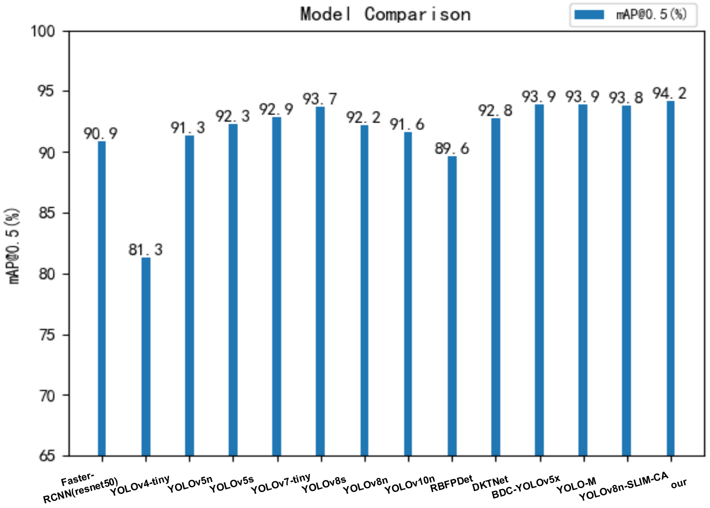
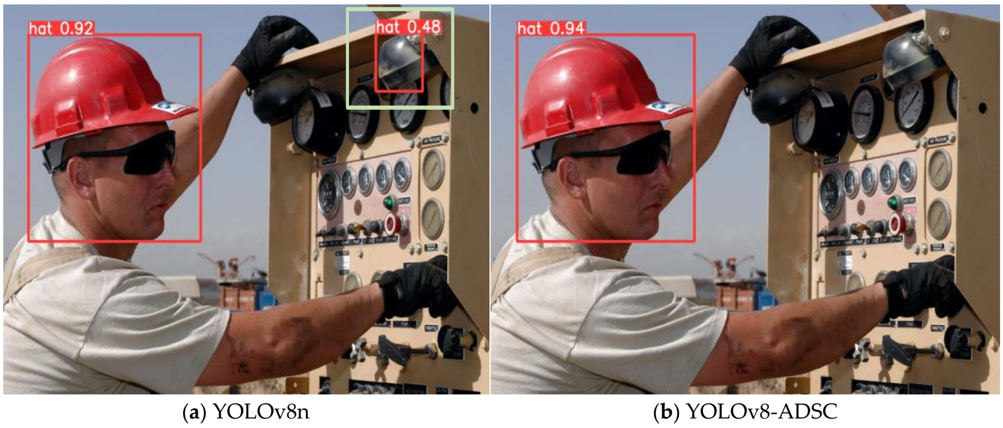


Figure 12. Comparison of mAP@0.5 values for all comparative experimental models.

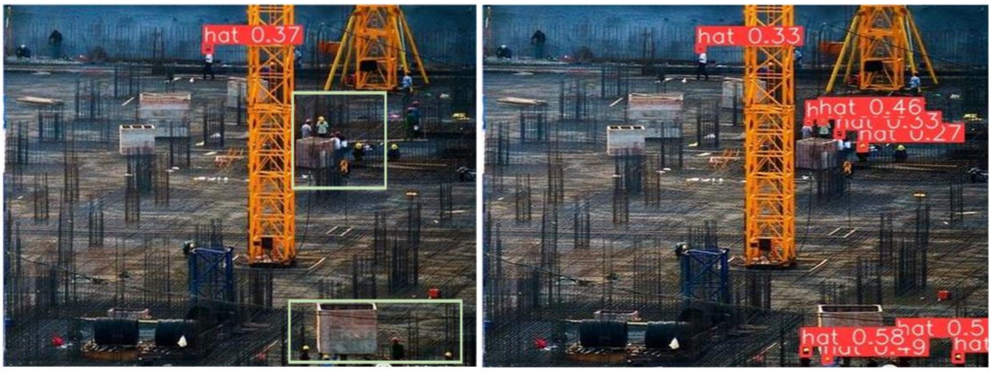
3.5. Visualization

This paper conducts a visual analysis under different working environments to provide a more intuitive illustration of YOLOv8-ADSC's effectiveness in detecting safety helmets within complex working environments. We select sample pictures from the test set's different working environments. In these complex working environments, the original YOLOv8n model has missed detection and false positives, and YOLOv8-ADSC has excellent detection effects on targets of different sizes and dense targets. The YOLOv8-ADSC model exhibits a notable enhancement in detection performance compared to the original YOLOv8n model. The test results are illustrated in Figures 13–15. The green box in the figure highlights changes in the test results.



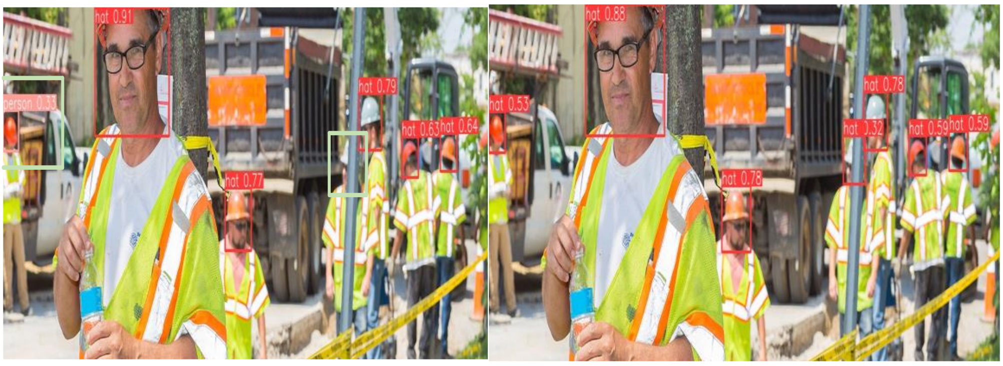
(**a**) YOLOv8n (**b**) YOLOv8-ADSC

Figure 13. Comparison of detection results of YOLOv8n and YOLOv8-ADSC under different types of interference.



(**a**) YOLOv8n (**b**) YOLOv8-ADSC

Figure 14. Comparison of detection results of far and small targets in gloomy light by YOLOv8n and YOLOv8-ADSC.



(**a**) YOLOv8n (**b**) YOLOv8-ADSC

Figure 15. Comparison of detection results of YOLOv8n and YOLOv8-ADSC in dense target environments.

The original model will have false detection phenomena under the interference of targets with similar shapes, miss detection phenomena for small targets at a distance under gloomy light, confuse positive samples with negative samples, and miss detection of blocked targets under a common dense target environment. Figures 13–15 visually show specific examples of this performance, showing that the improved YOLOv8-ADSC model effectively reduces the errors of the original YOLOv8n model, indicating that its detection effect is significantly improved.

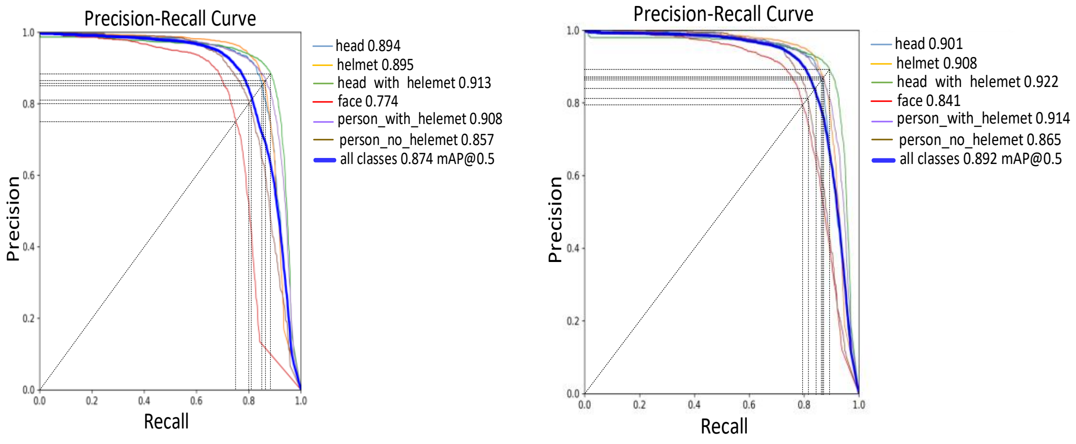
3.6. Generalization Experiment

This paper chooses to conduct generalization experiments on another safety helmet dataset, SHEL5K [31], as well as the public Unmanned Aerial Vehicle (UAV) dense small-target dataset, VisDrone2019 [32], to test the generalization capability of the YOLOv8-ADSC model proposed.

3.6.1. Generalization Experiments on the SHEL5K Dataset

SHEL5K is an improved Safety Helmet Detection [33] dataset. The original Safety Helmet Detection dataset contains 5000 images in three categories (person, head, and helmet). Three new labels have been added to the improved dataset SHEL5K, expanding its scope to encompass six categories: helmet, head with helmet, person with helmet, head, person without helmet, and face. This comprehensive dataset contains images of safety helmets in various scenarios, covering different wearing styles, environments, and lighting conditions, with a total of 75,578 labels. We use 3000 images for training, 1000 for validation, and 1000 for testing.

The original YOLOv8n and YOLOv8-ADSC models are used to conduct experiments on the SHEL5K dataset using the same equipment and environment. Figure 16 compares the PR curves between YOLOv8-ADSC and the original YOLOv8n. Table 5 shows the experimental results compared with the other models.



(**a**) YOLOv8n (**b**) YOLOv8-ADSC

Figure 16. Comparison of PR curves between the original YOLOv8n model and YOLOv8-ADSC in generalization experiments.

Table 5. Experimental results of YOLOv8-ADSC with conventional models on the SHEL5K.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	GFLOPs	FPS	Parameters (M)
Faster-R-CNN(ResNet50+FPN)	88.5	75	85.5	53.2	180	21.3	42
YOLOv4-tiny	81.1	62	68.3	41.6	6.79	59.2	5.89
YOLOv5n	88.3	76.2	83.8	49.9	4.1	101.4	1.9
YOLOv5s	88.8	78.5	85.7	52.5	15.8	98.5	7.03
YOLOv7-tiny	89.7	81.9	88.5	52.6	13	55.4	6.01
YOLOv8s	88.7	83.2	89.4	57	28.4	90.9	11.13
YOLOv8n	88.4	81	87.4	55.5	8.2	111	3.01
YOLOv10n	86.9	78.9	86.7	54.1	8.2	102	2.7
YOLOv8-ADSC	89	82.9	89.2	56.2	13.9	63.7	5.27

The experimental results in Table 5 show that on the SHEL5K dataset, the YOLOv8-ADSC model substantially enhances its detection capabilities, and the mAP@0.5 increases by 1.8%. Compared with the other traditional models, it still shows better detection performance, reflecting that YOLOv8-ADSC performs well on different safety helmet datasets.

3.6.2. Generalization Experiments on the VisDrone2019 Dataset

The dataset is created in cooperation between Tianjin University and the AISKYEYE data mining team. The UAV takes it from different heights and angles, containing ten categories of detection targets and many small targets. In this experiment, the dataset is partitioned into a training set (containing 6471 images), a validation set (containing 548 images), and a test set (containing 1601 images).

Table 6 lists the comparative experiment results of YOLOv8-ADSC and the other models on the VisDrone2019 dataset. The outcomes show that YOLOv8-ADSC has the best comprehensive performance, which is 6.8% higher than the mAP@0.5 of YOLOv8n, verifying the good generalization of the YOLOv8-ADSC model.

Table 6. Experimental results of YOLOv8-ADSC with conventional models on VisDrone2019.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	GFLOPs	FPS	Parameters (M)
Faster-R-CNN(ResNet50+FPN)	31.3	27.2	26.9	17.5	180	18	42
YOLOv4-tiny	31.3	11	18.8	10.7	6.79	55.6	5.89
YOLOv5n	36.3	28.8	26.1	13	4.1	59.8	1.9
YOLOv5s	39.3	31.7	28.7	15.4	15.8	59.5	7.03
YOLOv7-tiny	41.7	36	30.5	15.8	13	22.72	6.01
YOLOv8s	46.5	35.3	34.1	19.8	28.4	66	11.13
YOLOv8n	39.5	29	27.4	15.5	8.2	76	3.01
YOLOv10n	39.8	30	27.7	15.5	8.2	80	2.7
YOLOv8-ADSC	46.6	35.7	34.2	19.7	13.9	54.3	5.27

4. Conclusions

As the most basic and vital protective tool, the safety helmet plays an irreplaceable role in high-risk working environments. This paper proposes an improved YOLOv8n safety helmet detection model, YOLOv8-ADSC, which aims to enhance the detection performance of wearing helmets in complex working environments to minimize potential safety risks. Firstly, this paper enhances the detection head of the YOLOv8-ADSC model by integrating ASFF with DCNv2. We have substituted the conventional convolution in the detection module with DCNv2, improving the receptive field and effectively utilizing different scales' characteristics. This integration enables the model to capture key information more effectively in complex construction site backgrounds. Secondly, a new detection layer specifically designed for small-target detection is incorporated into the YOLOv8 model. This addition enhances the model's capability to identify and recognize small, remote targets with greater precision. The feature fusion mode of the ASFF module integrated with the detection head has also been updated to adapt to the four detection layers, further improving its detection performance. Finally, the original Upsample module is replaced with the lightweight CARAFE module. This substitution ensures that key feature information is retained during the up-sampling process, enhances the fusion of contextual information, and improves the model's capacity to utilize deep semantic information effectively. The ablation experiments verify the effectiveness of each improved part. The comparative experiments with traditional target detection models verify that the YOLOv8-ADSC model achieves optimal detection results, significantly lowering false and missed detections. The generalization experiments show that the YOLOv8-ADSC model has good generalization ability.

However, this model consumes more resources than the original model and is not efficient enough, which can be further improved by lightweight methods. Additionally, our current model primarily focuses on a single architecture, which may limit its ability to capture diverse semantic cues fully. In subsequent studies, we will use a combination of model pruning, target tracking, and other techniques while also exploring the integration of different network architectures to enhance semantic understanding. Our goal is to design a faster and more accurate safety helmet detection model that contributes more significantly to the field of safety production.

Author Contributions: Conceptualization, B.Z.; methodology, J.W.; software, B.S.; validation, J.W.; investigation, B.S. and W.L.; writing—original draft preparation, B.S. and J.W.; writing—review and editing, B.Z.; visualization, W.L.; supervision, B.Z. and J.W.; project administration, B.S.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Defense Industrial Technology Development Program under Grant JCKYS2022DC10 and the Foundation of Science and Technology Research Project of Colleges and Universities in Hebei Province under Grant QN2023140.

Data Availability Statement: The data are available on request due to privacy restrictions. The data presented in this study are available on request from the corresponding author.

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

References

1. Silva, R.R.V.E.; Aires, K.R.T.; Veras, R.D.M.S. Detection of helmets on motorcyclists. *Multimed. Tools Appl.* **2018**, *77*, 5659–5683. [CrossRef]
2. Sun, X.; Xu, K.; Wang, S.; Wu, C.; Zhang, W.; Wu, H. Detection and tracking of safety helmet in factory environment. *Meas. Sci. Technol.* **2021**, *32*, 105406. [CrossRef]
3. Gu, Y.; Wang, Y.; Shi, L.; Li, N.; Zhang, L.; Xu, S. Automatic detection of safety helmet wearing based on head region location. *IET Image Process.* **2021**, *15*, 2441–2453. [CrossRef]
4. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [CrossRef]
5. Lin, H.; Jiang, F.; Jiang, Y.; Luo, H.; Yao, J.; Liu, J. A model for Helmet-Wearing detection of Non-Motor drivers based on YOLOv5s. *Comput. Mater. Contin.* **2023**, *75*, 5321–5336. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
7. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–23 December 2015.
8. 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]
9. Lyu, Y.K.; Yang, X.; Guan, A.; Wang, J.; Dai, L. Construction personnel dress code detection based on YOLO framework. *CAAI Trans. Intell. Technol.* **2024**, *9*, 709–721. [CrossRef]
10. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
11. 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.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
13. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
14. 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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
15. Wang, M. Research on Detection and Application of Safety Helmet Wearing Based on Deep Learning. Master's Thesis, Anhui University of Technology, Maanshan, China, 2022.
16. Huang, L.; Fu, Q.; He, M.; Jiang, D.; Hao, Z. Detection algorithm of safety helmet wearing based on deep learning. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e6234. [CrossRef]
17. Xu, S.; Gu, j.; Hua, Y.; Liu, Y. DKTNet: Dual-Key Transformer Network for small object detection. *Neurocomputing* **2023**, *525*, 29–41. [CrossRef]
18. Wang, L.; Zhang, X.; Yang, H. Safety helmet wearing detection model based on improved YOLO-M. *IEEE Access* **2023**, *11*, 26247–26257. [CrossRef]
19. Zhao, L.; Tohti, T.; Hamdulla, A. BDC-YOLOv5: A helmet detection model employs improved YOLOv5. *Signal Image Video Process.* **2023**, *17*, 4435–4445. [CrossRef]
20. Lin, B. Safety helmet detection based on improved YOLOv8. *IEEE Access* **2024**, *12*, 28260–28272. [CrossRef]
21. Song, R.; Wang, Z. RBFPDet: An anchor-free helmet wearing detection method. *Appl. Intell.* **2023**, *53*, 5013–5028. [CrossRef]
22. Gao, T.; Yang, J.; Wang, W.; Fan, X. A domain feature decoupling network for rotating machinery fault diagnosis under unseen operating conditions. *Reliab. Eng. Syst. Saf.* **2024**, *252*, 110449. [CrossRef]
23. Gao, T.; Yang, J.; Tang, Q. A multi-source domain information fusion network for rotating machinery fault diagnosis under variable operating conditions. *Inf. Fusion* **2024**, *106*, 102278. [CrossRef]
24. Liu, Y.; Zhou, G.; Wu, S.; Han, J. TCGNet: Type-Correlation Guidance for Salient Object Detection. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 6633–6644. [CrossRef]
25. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
26. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
27. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
28. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-aware reassembly of features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
29. Amir, S.B.; Horio, K. YOLOv8s-NE: Enhancing Object Detection of Small Objects in Nursery Environments Based on Improved YOLOv8. *Electronics* **2024**, *13*, 3293. [CrossRef]
30. SHWD: Safety-Helmet-Wearing-Dataset. Available online: https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset (accessed on 1 June 2023).

31. Otgonbold, M.E.; Gochoo, M.; Alnajjar, F.; Ali, L.; Tan, T.H.; Hsieh, J.W.; Chen, P.Y. SHEL5K: An extended dataset and benchmarking for safety helmet detection. *Sensors* **2022**, *22*, 2315. [CrossRef] [PubMed]
32. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
33. Safety-Helmet-Detection: Improve Work Safety by Detecting the Presence of People and Safety Helmets. Available online: https://www.kaggle.com/andrewmvd/hard-hat-detection (accessed on 1 June 2023).

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.