electronics

MDPI

Article

A High-Precision Fall Detection Model Based on Dynamic Convolution in Complex Scenes

Yong Qin, Wuqing Miao * and Chen Qian

College of Measurement and Control Technology and Communication Engineering, Harbin University of Science and Technology, Harbin 150000, China; qinyong@hrbust.edu.cn (Y.Q.); 18845612770@163.com (C.Q.)
* Correspondence: m2717251033@163.com

Abstract: Falls can cause significant harm, and even death, to elderly individuals. Therefore, it is crucial to have a highly accurate fall detection model that can promptly detect and respond to changes in posture. The YOLOv8 model may not effectively address the challenges posed by deformation, different scale targets, and occlusion in complex scenes during human falls. This paper presented ESD-YOLO, a new high-precision fall detection model based on dynamic convolution that improves upon the YOLOv8 model. The C2f module in the backbone network was replaced with the C2Dv3 module to enhance the network's ability to capture complex details and deformations. The Neck section used the DyHead block to unify multiple attentional operations, enhancing the detection accuracy of targets at different scales and improving performance in cases of occlusion. Additionally, the algorithm proposed in this paper utilized the loss function EASlideloss to increase the model's focus on hard samples and solve the problem of sample imbalance. The experimental results demonstrated a 1.9% increase in precision, a 4.1% increase in recall, a 4.3% increase in mAP0.5, and a 2.8% increase in mAP0.5:0.95 compared to YOLOv8. Specifically, it has significantly improved the precision of human fall detection in complex scenes.

Keywords: complex scenarios; YOLOv8; deformable convolution; fall detection

Citation: Qin, Y.; Miao, W.; Qian, C. A High-Precision Fall Detection Model Based on Dynamic Convolution in Complex Scenes. *Electronics* **2024**, *13*, 1141. https://doi.org/10.3390/electronics13061141

Academic Editor: George A. Tsihrintzis

Received: 23 February 2024
Revised: 15 March 2024
Accepted: 16 March 2024
Published: 20 March 2024

1. Introduction

Individuals are susceptible to falls due to instability in their lower extremities and limited joint mobility during daily activities [1]. The likelihood and severity of falls are particularly high in individuals over the age of 65, with 30–40% experiencing at least one fall per year. These falls can result in fractures or other long-term health issues, which can cause significant physical and psychological injury [2–4]. Injuries sustained by older adults from falls depend not only on the injuries incurred but also on the time interval between the onset of the fall and the receipt of help and treatment. Medical research has shown that timely assistance or treatment after a fall can reduce the risk of sequelae from later falls as well as accidental death [5]. Providing timely assistance and treatment services for elderly individuals who live alone and have fallen at home is of significant social and practical importance. This ensures the safety and security of the elderly.

Currently, there are several methods for detecting human posture [6] which can also recognize and detect falls. Wearable technology development can integrate sensors, wireless communication, and other technologies into wearable devices. These devices support various interaction methods, such as gesture and eye movement, to capture human body movement and posture information. They use multi-information data fusion to achieve the detection of human falls, resulting in high detection accuracy and real-time detection [7–10]. However, older people may forget to wear them after charging, which hinders prolonged detection due to the need for frequent recharging. Placing sensor nodes in a specific area to monitor changes in the human body's center of gravity, movement trajectory, and position can provide valuable information about the body's posture and

overall situation [11–14]. However, deployment costs are high, and external environmental limitations and interference can be a challenge.

The utilization of cameras or other imaging devices for real-time acquisition of image information in a monitoring area, coupled with the application of deep learning techniques to analyze the acquired image data and determine human body movement postures, represents a current research focus [15,16]. Deep learning methods for analysis can be broadly categorized into two directions: two-stage and one-stage [17]. Prominent examples of two-stage algorithms include R-CNN, Mask R-CNN [18], R-FCN [19], and Faster R-CNN [20]. These approaches offer advantages such as high detection rates and low memory usage [21,22]. On the other hand, one-stage algorithms like the YOLO series [23–25] and the SSD series perform candidate frame generation and classification in a single step. By dividing images into grids, these algorithms directly predict target categories and anchor frames on the images before obtaining final results through filtering and post-processing. Due to their lower computational requirements, one-stage algorithms are more suitable for real-time detection projects. Furthermore, recent advancements in the YOLO series have significantly improved target detection accuracy, establishing the one-stage algorithm as the mainstream choice for practical applications. Therefore, this paper selects YOLOv8 as its foundation to enhance fall detection in complex scenes.

The YOLOv8 model represents the latest advancement in the YOLO series, incorporating novel enhancements derived from YOLOv5 to optimize performance and flexibility, thereby rendering it more suitable for diverse target detection tasks. Comprising three key components—the backbone, neck network, and detection head—this model leverages the C2f module within its backbone network to effectively merge the C3 and ELAH structures, facilitating superior feature transfer and enhancing information utilization efficiency. Notably, the YOLOv8 detection head adopts a decoupled head approach by eliminating the objectness branch while retaining only classification and regression branches. This simplification significantly streamlines the model architecture. Additionally, an Anchor Free strategy is employed which eliminates reliance on predefined anchors; instead enabling adaptive learning of object size and position. Consequently, these advancements contribute to improved accuracy and robustness in object detection.

Lijuan Zhang et al. proposed DCF-YOLOv8, which leverages DenseBlock to enhance the C2f module and mitigate the influence of environmental factors [26]. Haitong Lou et al. introduced DC-YOLOv8, employing deeper networks for the precise detection of small targets [27]. Gui Xiangquan et al. incorporated the DepthSepConv lightweight convolution module into YOLOv8-L, integrated the BiFormer attention mechanism, and expanded the small target detection layer to achieve efficient detection of small targets [28]. Cao Yiqi et al., in EFD-YOLO, substituted EfficientRep as a backbone network and introduced the FocalNeXt focus module to address occlusion issues to some extent while enhancing detection accuracy [29].

To address the issue of low detection accuracy of the YOLOv8 algorithm in complex environments with target deformation, large changes in target scale, and occlusion, this paper proposes the ESD-YOLO model based on the YOLO algorithm. The model incorporates dynamic convolution, a dynamic detection head, and an exponential moving average to enhance the accuracy and robustness of fall detection in complex scenarios. This paper presents research on improving the YOLOv8 backbone network's ability to capture target details and cope with target deformations. The proposed C2Dv3 module was incorporated into the network for this purpose. Additionally, the feature extraction ability of the detection model was improved by replacing the original detection head in the Neck section with the DyHead module. The proposed EASlideloss loss function aims to improve the model's ability to handle hard sample problems. ESD-YOLO performed better in dim and blurred environments, with informative pictures, large-scale transformations, and occlusions, improving the accuracy and robustness of the fall detection model.

2. Materials and Methods

2.1. Overall Structure of ESD-YOLO Network

This paper proposed ESD-YOLO, a high-precision fall detection model for complex scenarios. It effectively addressed the problem of low detection accuracy caused by fall target deformation, occlusion, and high environmental overlap. Figure 1 shows the overall structural model of ESD-YOLO.

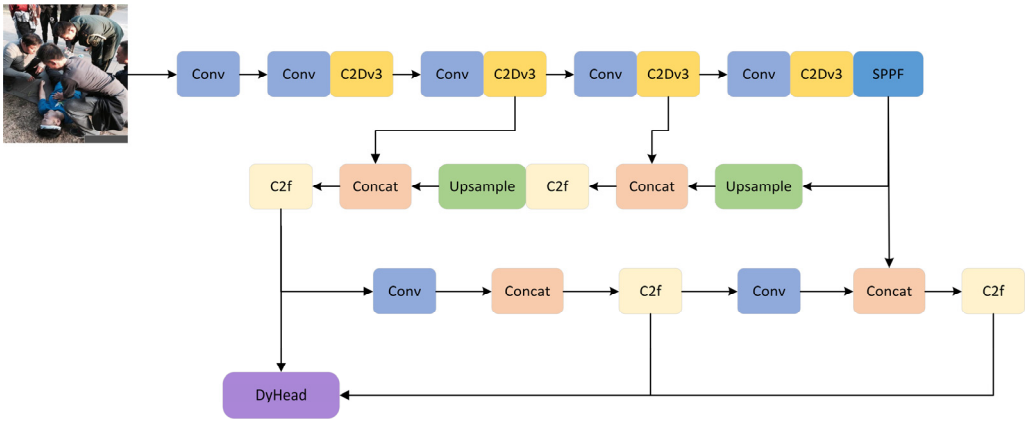

Figure 1. Structure of ESD-YOLO model.

The ESD-YOLO model combined the C2f module in the YOLOv8 backbone with the DCNv3 module. The dynamic convolutional layer replaced the convolutional layer in the Bottleneck in C2f, enhancing the backbone network's ability to extract pose information of a falling character in a complex scene. The DyHead module was incorporated into the Neck section to consolidate multiple attention operations, resulting in improved performance of ESD-YOLO in complex fall detection scenarios. Additionally, EASlideloss, a slide loss function based on exponential moving average, was proposed to replace the original loss function of YOLOv8. This function dynamically balances the model's attention to hard samples, thereby enhancing the model's accuracy and stability.

2.2. C2Dv3 Module Design

Detecting falls in complex environments and a wide variety of poses presents a significant challenge. The C2f module in YOLOv8, which integrates low-level feature maps with high-level feature maps, encounters difficulties in recognizing falls under these circumstances. The C2f module may not effectively capture the intricate details of falling targets due to variations in human body postures, resulting in substantial changes in target size and shape. Moreover, the module is limited to sensing features within a fixed range and lacks the adaptability to adjust the sampling position of the convolution kernel dynamically, making it arduous to capture crucial information about falling targets comprehensively. Consequently, this led to decreased accuracy for target localization in complex environments and increased the likelihood of false detections.

To address the limitations of the C2f module in detecting falls with significant variations in scale and high environmental similarity, we introduced DCNv3 during the feature extraction stage. DCNv3 effectively captures comprehensive information surrounding the fall target within the sensory field and adapts to diverse sizes and shapes by dynamically adjusting convolution kernel shapes and positions [30]. The deformable convolution operation in DCNv3 employs a learnable offset to govern the shape of each convolution kernel, thereby facilitating adaptive adjustment of the convolution operation based on diverse image regions and enhancing its perceptual capability. This enhancement enables a more precise capture of fall target details and features, thereby improving both the accuracy and robustness of our fall detection model. Consequently, it led to enhanced precision in detecting fall targets and increased reliability of the model even in complex scenarios.

The DCNv3 model enables adaptive modification of the convolution kernel shape based on the target content in the image. This flexible mapping enhances the coverage of the detected target appearance and captures a more comprehensive range of useful feature information [30]. Equation (1) represents the expression for DCNv3.

$$y(p_0) = \sum_{g=1}^{G} \sum_{k=1}^{K} w_g m_{gk} x_g \left(p_0 + p_k + \Delta p_{gk} \right) \quad (1)$$

Equation (1) defines G as the number of groups, w_g as the projection weights shared within each group, and m_{gk} as the normalized modulation factor of the Kth sampling point of the Gth group. DCNv3 exhibits superior adaptability to large-scale visual models compared to its counterparts in the same series, while also possessing stronger feature representation and a more stable training process.

DCNv3 has negligible impact on the number of parameters or computational complexity of the model. However, excessive utilization of deformable convolutional layers can significantly increase computation time in practical applications. To ensure optimal performance without compromising functionality, only the standard convolutional layers within the Bottleneck of the C2f module in the backbone network were substituted with DCNv3 deformable convolutional layers, forming a compliant bottleneck module (Dv3_Bottleneck), as depicted in Figure 2.

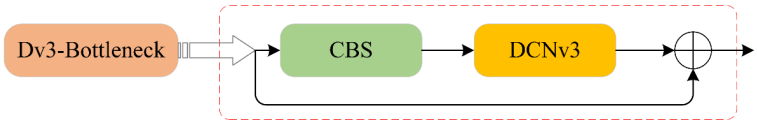

Figure 2. DCNv3 replaces Standard Conv.

As illustrated in Figure 3, the C2f module has been reconstructed using Dv3_Bottleneck, which comprises of convolution layer, separation layer, and Dv3_Bottleneck. The incorporation of C2Dv3 into the backbone network enhances its ability to capture crucial target features, thereby elevating target detection performance.

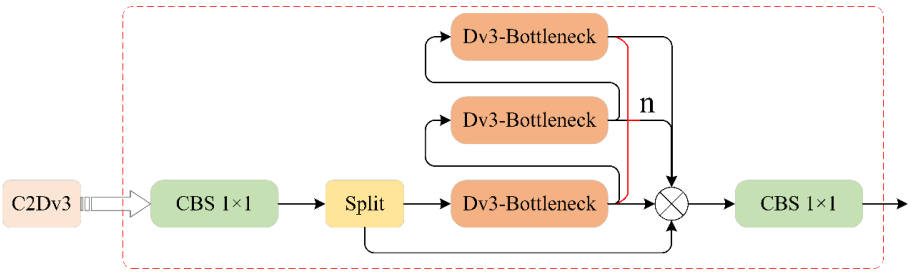

Figure 3. C2Dv3 model based on Dv3-Bottleneck.

2.3. DyHead Module

To better integrate the diversity of feature scales resulting from variations in falling target scale and capture the inherent spatial relationships across different scales and shapes, this study replaced the original detection head of YOLOv8 with a dynamic detection head called DyHead (Dynamic Head). DyHead incorporates scale-aware attention, spatial-aware attention, and task-aware attention simultaneously [31]. It employs a dynamic receptive field design that adaptively adjusts the convolution kernel size based on the size of the falling target. The DyHead model possesses the capability to integrate multiple attention mechanisms, thereby enabling the fusion of diverse information and mitigating the adverse effects caused by occlusion. This ensures effective detection of targets with varying scales

and shapes while enhancing overall detection capability and optimizing computational efficiency. The calculation formula is presented in Equation (2).

$$W(F) = \pi_C(\pi_S(\pi_L(F)\cdot F)\cdot F)\cdot F \tag{2}$$

The attention function is represented by the symbol W, and the feature tensor F is a three-dimensional tensor with dimensions of $L \times S \times C$. Here, L represents the level of the feature map, S represents the width-height product of the feature map, and C represents the number of channels in the feature map. The scale-aware attention module $\pi_L(\cdot)$, space-aware attention module $\pi_S(\cdot)$, and task-aware attention module $\pi_C(\cdot)$ are, respectively, applied to each dimension of L, S, and C. Figure 4 illustrates the structure of a single DyHead block.

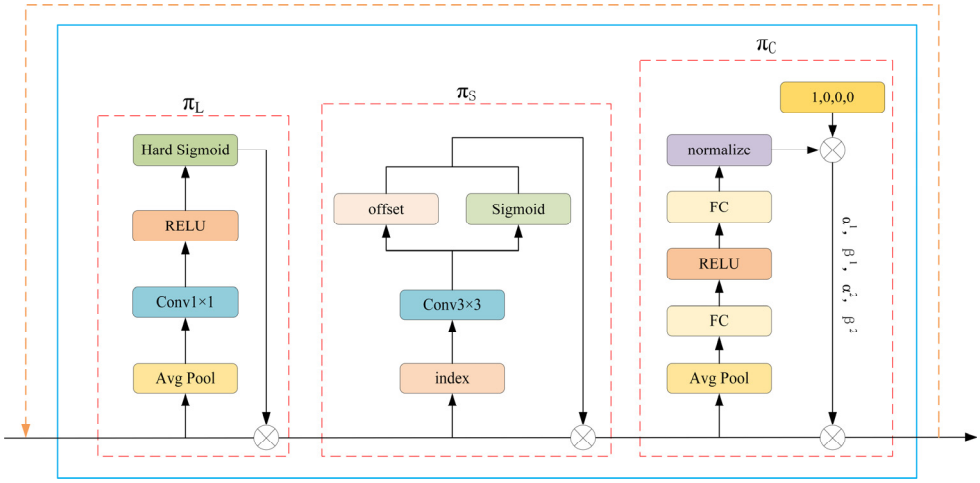

Figure 4. Structure of DyHead model.

The computational processes for each of the three attention modules are represented as follows:

$$\pi_L(F)\cdot F = \sigma\left(f\left(\frac{1}{SC}\sum_{S,C}F\right)\right)\cdot F \tag{3}$$

$$\pi_S(F)\cdot F = \frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{K} w_{l,j}\cdot F\left(l; p_j + \Delta p_j; c\right)\cdot \Delta m_j \tag{4}$$

$$\pi_C(F)\cdot F = \max\left(\alpha^1(F)\cdot F_C + \beta^1(F), \alpha^2(F)\cdot F_C + \beta^2(F)\right) \tag{5}$$

In Equation (3), the linear function $f(\cdot)$ is approximated using a 1×1 convolution operation. Herein, $\sigma(x) = max(0, min(1, (x+1)/2))$ serves as an activation function for this approximation process. Before introducing K as representing the number of sparse sampling positions in Equation (4), we explain that these positions enable focusing on discriminant locations through determining movable position $p_j + \Delta p_j$ based on self-learning spatial displacement Δp_j. Moreover, we introduce Δm_j, denoting a self-learning importance scalar at position p_j which can be learned from input features at middle level F. Subsequently defined in Equation (5), F_C refers to the feature slice of channel C while $\left[\alpha^1, \beta^1, \alpha^2, \beta^2\right]^T = \theta(\cdot)$ represents a superfunction employed for learning control activation threshold values. Sequentially applying these three attention mechanisms allows them to be stacked multiple times to form DyHead blocks, as depicted in Figure 5.

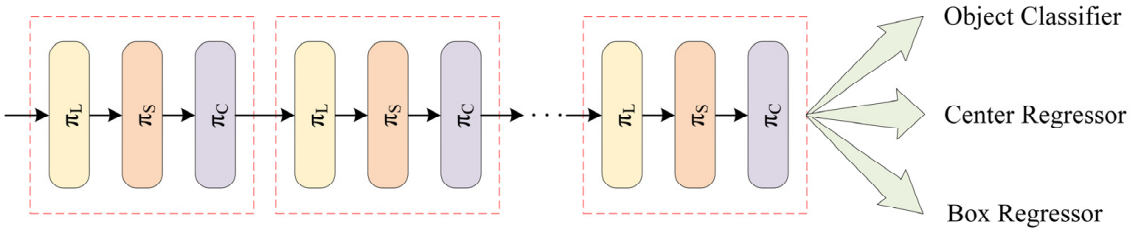

Figure 5. Connection scheme of DyHead blocks.

2.4. Loss Function EASlideloss Design

The elderly population is more susceptible to falls in complex environments, and the fall detection model encounters challenges such as obscured fall objects, low ambient lighting, high environmental overlap, and diverse fall postures. These data present hard samples with a lower number of fall instances compared to non-fall instances, resulting in an imbalanced dataset. Without an appropriate loss function, the performance of the fall detection model in the target category is compromised, thereby affecting its accuracy and reliability in practical applications. YOLOv8's original BCEwithloss (BCE) loss function solely focuses on accurate label prediction without addressing sample balancing when tackling the sample imbalance issue. Consequently, the model prioritizes non-fall instances over effectively identifying falling actions. To address this limitation, Slideloss incorporates a sliding window mechanism that adaptively learns threshold parameter μ for positive and negative samples. By assigning higher weights near μ, it amplifies relative loss for hard-classified samples while emphasizing misclassified ones [32]. The implementation principle is illustrated by Equation (6).

$$f(x) = \begin{cases} 1 & x \leq \mu - 0.1 \\ e^{(1-\mu)} & \mu - 0.1 < x < \mu \\ e^{(1-x)} & x \geq \mu \end{cases} \tag{6}$$

The proposed EASlideloss in this paper is based on Slideloss, which integrates the exponential moving average (EMA) with the original Slideloss. By applying the exponential moving average method to weigh the value of the time series, we aim to mitigate the impact of sudden changes in adaptive threshold on loss and enhance both the accuracy and reliability of our model. Additionally, we gradually reduce the weight assigned to difficult samples, thereby diminishing the model's attention towards them and preventing excessive interference caused by these challenging instances throughout the training process. The implementation principle is illustrated in Equations (7)–(9).

$$\mathrm{d}_i = \beta\left(1 - e^{-\frac{i}{\tau}}\right) \tag{7}$$

$$\mu = \mathrm{d}_i \cdot \mu_{i-1} + (1 - d_i) \cdot \theta_i \tag{8}$$

$$f(x) = \begin{cases} 1 & x \leq \mu - 0.1 \\ e^{(1-\mu)} & \mu - 0.1 < x < \mu \\ e^{(1-x)} & x \geq \mu \end{cases} \tag{9}$$

In Equation (7), the attenuation factor $0 < d_i < 1$ represents the weight distribution control for historical and latest data when calculating the average value, where β denotes the attenuation coefficient. The variable i represents the current training round, while τ is a hyperparameter. In Equation (8), μ_{i-1} signifies the previous time's average index value, and θ_i represents the current time's data.

2.5. Model Evaluation Metrics

The evaluation metrics employed in this study to assess the performance of the fall detection model include Precision (P), Recall (R), and Average Precision (AP). AP quantifies the detector's performance within each category, while the mean average precision (mAP) is obtained by averaging these AP values. mAP serves as a pivotal metric for evaluating the overall accuracy of object detection models and a reliable indicator of their performance.

3. Experiment and Results

3.1. Datasets

Fall events are relatively uncommon in daily life. Although existing public fall detection datasets attempt to simulate the complex and authentic nature of falls, they still suffer from limitations such as simplistic experimental environments and an inability to accurately replicate real-life falls. In this study, we comprehensively utilized the UR Fall Detection Dataset, the Fall Detection Dataset, and images of human falls collected from real-world scenes on the Internet to gather data encompassing different illuminations, angles, object similarities, and occlusion scenarios. A total of 4976 datasets were obtained through this process. Subsequently, we employed the open-source tool LabelImg to uniformly label these data images in Yolo format and generate corresponding labels for a total of 5655 samples depicting various poses. The dataset was then divided into a training set (70%), a test set (20%), and a validation set (10%) following a 7:2:1 ratio format. Figure 6 illustrates some representative scenes from our dataset that can serve as references for evaluating the performance of ESDv3-YOLO under realistic conditions.

Figure 6. Typical dataset presentation.

3.2. Experimental Process

The test platform is configured with a 6-core E5-2680 v4 processor and an NVIDIA GeForce RTX 3060 GPU. The operating system used is Windows 11, along with PyTorch version 2.0.1 in the development environment of PyCharm 2022.2.3 and Python version 3.10.12. The model takes input images of size 640 × 640 pixels for training purposes, while the training parameters consist of a batch size of 32, a total of 200 iterations, momentum set to 0.937, initial learning rate at 0.001, and an attenuation coefficient value of 0.9.

The YOLOv8s and ESD-YOLO models share the same dataset and training parameters, as depicted in Figure 7. Following 20 iterations, both models exhibit a gradual decline in loss value, with the error reaching stability after 75 iterations. Experimental findings demonstrate that compared to the original model, ESD-YOLO showcases accelerated convergence speed, reduced loss value, and significantly enhanced network convergence capability.

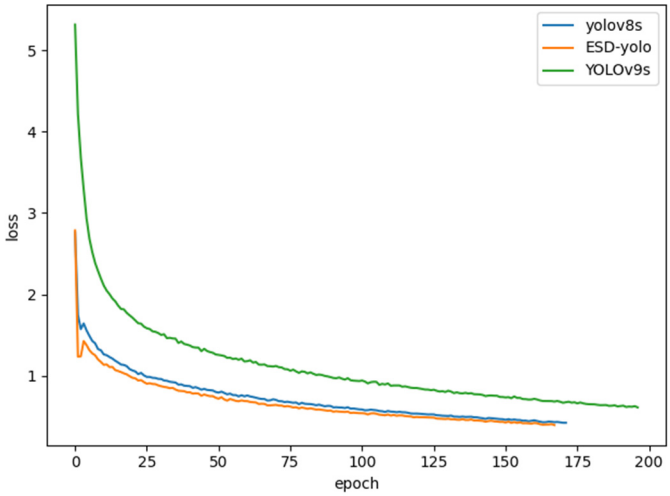

Figure 7. Loss curves are trained by three models.

3.3. Experimental Results and Analysis

To validate the performance of the ESD-YOLO model, we conducted two sets of comparative experiments. The first set aimed to compare the accuracy and performance of the improved fall detection model with YOLOv8s. The second set further compared the differences between the improved model and both YOLO series models and mainstream object detection algorithms. Through these comprehensive comparisons, we can thoroughly evaluate the accuracy and performance of our improved fall detection model in comparison to other relevant algorithms.

3.3.1. Ablation Experiment

The ablation experiment aims to validate the optimization effect of each enhanced module. In this study, we conducted an ablation analysis on ESD-YOLO, where specific enhancements were incorporated into the YOLOv8s model, namely C2Dv3, DyHead, and EASlideloss denoted as YOLOv8s_1, YOLOv8s_2, and YOLOv8s_3. As depicted in Table 1, each module exhibited varying degrees of improvement in the accuracy of ESD-YOLO.

Table 1. Ablation experiments with different design strategies.

Modules	C2Dv3	DyHead	EASlideloss	P (%)	R (%)	mAP0.5 (%)	mAP0.5:0.95 (%)
YOLOv8s				82.3	78.4	84.4	59.7
YOLOv8s_1	√			84.7	80.2	86.4	62.5
YOLOv8s_2		√		85.9	78.2	86.1	61.8
YOLOv8s_3			√	76.8	83.7	84	60
ESD-YOLO	√	√	√	84.2	82.5	88.7	62.5

After incorporating the C2Dv3 module to enhance the backbone network, there is a noticeable improvement in precision (2.4%), recall (1.8%), mAP0.5 (2%), and mAP0.5:0.95 (2.8%). These findings demonstrate that the C2Dv3 module effectively enhances the feature extraction capability of the backbone network, enabling it to accurately capture intricate details of falling human bodies and effectively handle target deformations. Moreover, this module exhibits superior accuracy in recognizing positive samples and enhances its ability to identify genuine positive instances, thereby enhancing overall detection performance.

After incorporating DyHead into the Neck component, the accuracy, mAP0.5, and mAP0.5:0.95 witnessed a respective increase of 3.6%, 1.7%, and 2.1%. This observation substantiates that replacing the original detection head of YOLOv8 with DyHead effectively enhances adaptability towards scale transformations and shape variations in detected objects, thereby augmenting model perception ability and accuracy.

By replacing the original BCEwithloss function in YOLOv8 with EASlideloss, a significant improvement of 5.3% in recall rate and 0.3% in mAP0.5:0.95 was observed, indicating that the utilization of EASlideloss enhances fall detection accuracy and reliability, thereby augmenting the model's ability to accurately detect falls.

The results of the ablation experiment demonstrate that all three enhanced modules contribute to improved accuracy of the overall model, indicating a strong coupling between these refined methods. Consequently, ESDv3-YOLO exhibits a significant performance enhancement in comparison with the original YOLOv8s.

3.3.2. Contrast Experiment

In order to assess the accuracy of various high-performance models for fall detection, we selected 11 representative network models, namely YOLOv4-tiny, YOLOv5s, YOLOv5-timm, YOLOv5-efficientViT, YOLOv5-vanillanet, YOLOv7, YOLOv7-tiny, YOLOv8s, YOLOv9s, SSD, and Faster R-CNN for comparative testing with ESD-YOLO. All models were trained and tested using the same dataset.

The fall detection results of different models are presented in Table 2. It is observed that the ESD-YOLO model achieves the highest accuracy, mAP0.5, and mAP0.5:0.95 values among the aforementioned 11 models, with respective scores of 84.2%, 88.7, and 62.5%. In comparison to these 11 network models, the ESD-YOLO model demonstrates improvements in mAP0.5 by 10.2%, 3.2%, 6.7%, 4.4%, 5.5%, 10.8%, 3.2%, 6.8%, 4.3%, 2%, 12.6%, and 7.9%. In addition, Map0.5:0.95 improves by 6.9%, 2.6%, 3%, 3.9%, 5.6%, 2.9%, 2.1%, 4.2%, 2.8%, 1.1%, 8.6%, and 5.8%, respectively. The ESD-YOLO algorithm exhibits significant performance advantages when compared to mainstream algorithms due to its comprehensive consideration of spatial transformation and shape information, enabling it to perform well even under conditions involving large-scale transformations and occlusions of falling targets. Therefore, in contrast to other algorithms, ESD-YOLO demonstrates superior adaptability for fall detection tasks.

Table 2. Comparative experiments on fall detection results of high-precision models.

Modules	P (%)	R (%)	Map0.5 (%)	Map0.5:0.95 (%)
YOLOv4-tiny	75.9	77.4	78.5	55.6
YOLOv5s	82.3	79.9	85.5	59.9
YOLO5-timm	81.2	78.9	82	59.5
YOLOv5-efficientViT	83.1	78.5	84.3	58.6
YOLOv5-vanillanet	78.3	77.5	83.2	56.9
YOLOv5-ShuffleNetv2	78.1	**83.1**	77.9	59.6
YOLOv7	80.2	80.6	85.5	60.4
YOLOv7-tiny	78.2	82.2	81.9	58.3
YOLOv8s	82.3	78.4	84.4	59.7
YOLOv9s	**84.3**	79.2	86.7	61.4
SSD	76.2	71.8	76.1	53.9
Faster R-CNN	80.7	77.8	80.8	56.7
ESDv3-YOLO	84.2	82.5	**88.7**	**62.5**

3.4. Scene Test

A comprehensive visual comparison between YOLOv8s and ESD-YOLO algorithms was conducted in various scenarios, encompassing scenes with significant scale changes of falling targets, dense crowds, high environmental similarity, and target occlusion. The detailed comparison is presented in Figure 8.

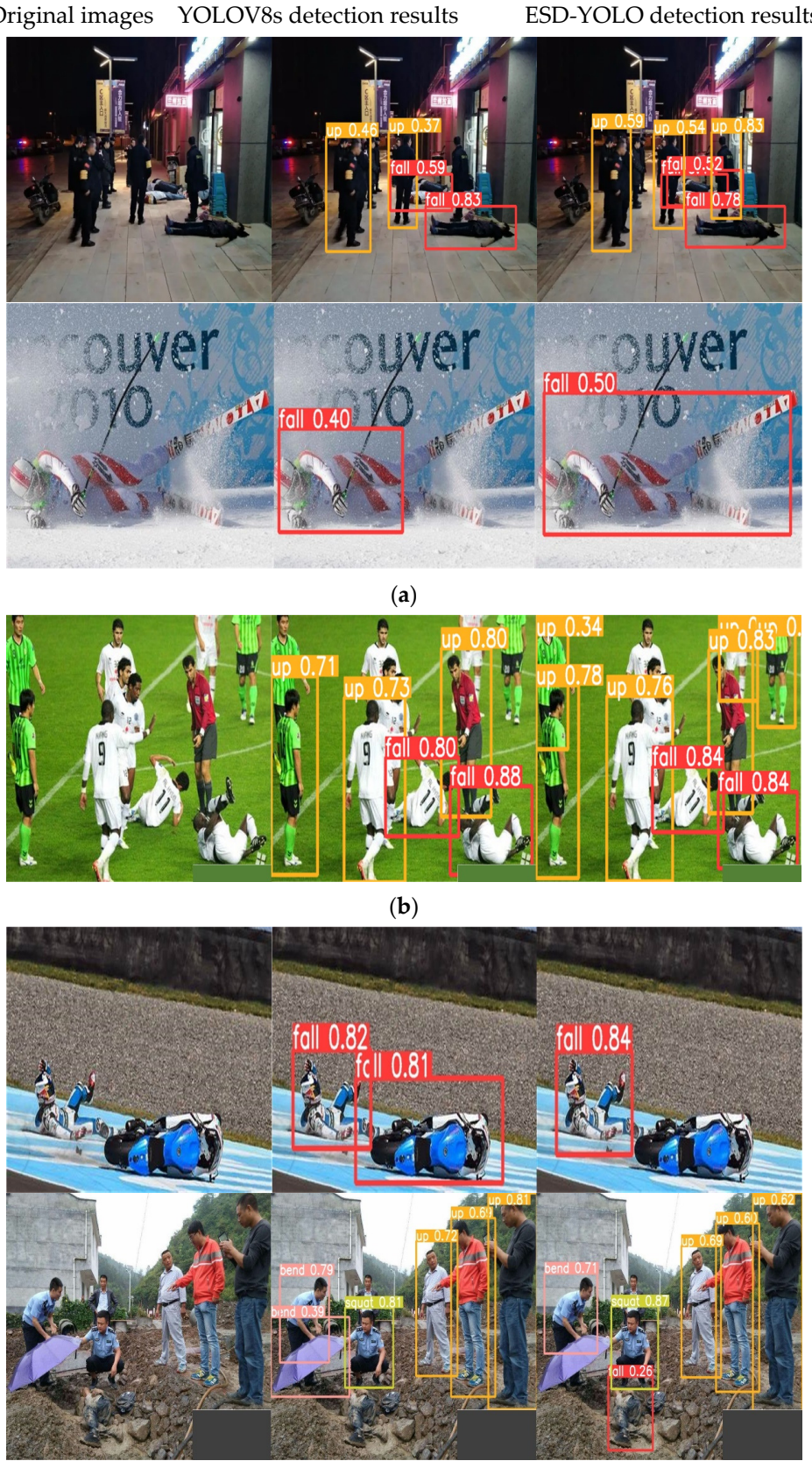

Figure 8. *Cont*.

(**d**)

Figure 8. Test in real scenarios. (**a**) Detection targets with large-scale variations; (**b**) intensive fall detection targets; (**c**) identify difficult scenes; (**d**) target occlusion.

In Figure 8a, the person's size in the image is either too large or too small, resulting in missed detections and incorrect box selections. In comparison to YOLOv8s, ESD-YOLO demonstrates improved capability in identifying more detection targets and accurately selecting boxes with higher confidence. This improvement can be attributed to the integration of the C2Dv3 module, which enhances network receptive field and feature extraction abilities. Consequently, ESD-YOLO effectively focuses on detecting targets, significantly enhancing prediction box positioning accuracy while minimizing missed detections. As a result, it achieves superior prediction accuracy for fall detection tasks.

The people in Figure 8b are densely packed with complex spatial positions, resulting in missed detection by YOLOv8s. By replacing DyHead as the detection head, ESD-YOLO can better capture spatial information and identify more monitoring targets than YOLOv8s, while also exhibiting higher confidence in detecting falling figures.

In Figure 8c, there exist objects resembling the falling target, and due to a significant overlap between the falling target and its surroundings, fall detection becomes considerably more challenging. By incorporating EASlideloss and C2Dv3 into ESD-YOLO, our approach effectively focuses on difficult samples and captures crucial information regarding the relationship between the falling target and its environment. This leads to a reduced probability of false detection and improved accuracy in detecting falls.

The falling target in Figure 8d is evidently obstructed, leading to a failure of YOLOv8s in identifying the target and resulting in false detection. In contrast, ESD-YOLO places greater emphasis on challenging samples and effectively addresses the issue of difficult identification caused by blockage by leveraging spatial position information encompassing the detection target's surroundings.

Based on the aforementioned experimental analysis, it is evident that ESD-YOLO exhibits superior performance in intricate environments.

4. Conclusions

The present study introduces ESD-YOLO, a high-precision algorithm for human fall detection in complex scenes. In comparison to the YOLOv8s model, it exhibits enhanced capabilities in addressing challenges encountered during fall tasks, including large target scale transformations, crowded environments with multiple individuals, and high levels

of environmental fusion and occlusion. The main contributions of this paper can be summarized as follows:

The C2Dv3 module is proposed to redesign the backbone network of YOLOv8s, enhancing its feature extraction ability and enabling it to better capture details of falling human bodies and process complex features of falling targets.

DyHead replaces the original detection head of YOLOv8s, allowing the model to focus on potential position relationship features of falling targets in different scales and shapes in spatial positions.

EASlideloss loss function replaces the original BCE loss function of YOLOv8s, improving accuracy while ensuring stability by focusing on difficult fall samples and gradually reducing attention to them.

The experimental results on the self-constructed dataset demonstrate that ESD-YOLO achieves an accuracy of 84.2%, a recall of 82.5%, a mAP0.5 of 88.7%, and a mAP0.5:0.95 of 62.5%. In comparison with the original YOLOv8s model, ESD-YOLO exhibits improvements in accuracy, recall, mAP0.5, and mAP0.5:0.95 by 1.9%, 4.1%, 4.3%, and 2.8%, respectively. The comprehensive fall detection experiments validate that ESD-YOLO possesses an efficient architecture and superior detection accuracy, thereby meeting the real-time fall detection requirements effectively. Furthermore, when compared to existing fall detection models, ESD-YOLO offers enhanced detection accuracy for various complex fall scenarios. In summary, ESD-YOLO enhances the accuracy of human fall detection and enables real-time identification and alerting of falls. It facilitates timely detection of elderly individuals experiencing falls and transmits alarm information to their caregivers through various communication channels, thereby enabling prompt intervention. Future research directions should focus on reducing model parameters to facilitate its deployment on mobile devices, making it applicable in real-world scenarios.

Author Contributions: Conceptualization, W.M. and C.Q.; Data curation, W.M. and C.Q.; Formal analysis, Y.Q. and W.M.; Funding acquisition, Y.Q.; Investigation, W.M., Y.Q. and C.Q.; Project administration, Y.Q.; Resources, Y.Q., W.M. and C.Q.; Writing—original draft, W.M.; Writing—review and editing, W.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in my paper are all publicly available datasets, which are composed of three datasets in total. Here are the links to the three publicly available datasets: (https://opendatalab.com/BoosCrob/FallDet1000 (accessed on 15 September 2023), http://fenix.ur.edu.pl/~mkepski/ds/uf.html (accessed on 15 September 2023), and https://falldataset.com/ (accessed on 15 September 2023)). After downloading them, I made datasets suitable for my use through my own annotations. 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. Lin, Y.-T.; Lee, H.-J. Comparison of the Lower Extremity Kinematics and Center of Mass Variations in Sit-to-Stand and Stand-to-Sit Movements of Older Fallers and Nonfallers. *Arch. Rehabil. Res. Clin. Transl.* **2022**, *4*, 100181. [CrossRef]
2. Gates, S.; Fisher, J.; Cooke, M.; Carter, Y.; Lamb, S. Multifactorial assessment and targeted intervention for preventing falls and injuries among older people in community and emergency care settings: Systematic review and meta-analysis. *BMJ* **2008**, *336*, 130–133. [CrossRef]
3. Pozaic, T.; Lindemann, U.; Grebe, A.-K.; Stork, W. Sit-to-Stand Transition Reveals Acute Fall Risk in Activities of Daily Living. *IEEE J. Transl. Eng. Healthc. Med.* **2016**, *4*, 2700211. [CrossRef]
4. Xu, T.; An, D.; Jia, Y.; Yue, Y. A Review: Point Cloud-Based 3D Human Joints Estimation. *Sensors* **2021**, *21*, 1684. [CrossRef]
5. Wang, X.; Ellul, J.; Azzopardi, G. Elderly fall detection systems: A literature survey. *Front. Robot. AI* **2020**, *7*, 71. [CrossRef] [PubMed]
6. Dai, Y.; Liu, W. GL-YOLO-Lite: A Novel Lightweight Fallen Person Detection Model. *Entropy* **2023**, *25*, 587. [CrossRef] [PubMed]
7. Wang, S.; Miranda, F.; Wang, Y.; Rasheed, R.; Bhatt, T. Near-Fall Detection in Unexpected Slips during Over-Ground Locomotion with Body-Worn Sensors among Older Adults. *Sensors* **2022**, *22*, 3334. [CrossRef] [PubMed]

8. Chander, H.; Burch, R.F.; Talegaonkar, P.; Saucier, D.; Luczak, T.; Ball, J.E.; Turner, A.; Kodithuwakku Arachchige, S.N.K.; Carroll, W.; Smith, B.K.; et al. Wearable Stretch Sensors for Human Movement Monitoring and Fall Detection in Ergonomics. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3554. [CrossRef] [PubMed]
9. Lee, Y.; Pokharel, S.; Muslim, A.A.; Kc, D.B.; Lee, K.H.; Yeo, W.-H. Experimental Study: Deep Learning-Based Fall Monitoring among Older Adults with Skin-Wearable Electronics. *Sensors* **2023**, *23*, 3983. [CrossRef] [PubMed]
10. Er, P.V.; Tan, K.K. Wearable solution for robust fall detection. In *Assistive Technology for the Elderly*; Academic Press: Cambridge, MA, USA, 2020; pp. 81–105.
11. Bhattacharya, A.; Vaughan, R. Deep learning radar design for breathing and fall detection. *IEEE Sens. J.* **2020**, *20*, 5072–5085. [CrossRef]
12. Jiang, X.; Zhang, L.; Li, L. Multi-Task Learning Radar Transformer (MLRT): A Personal Identification and Fall Detection Network Based on IR-UWB Radar. *Sensors* **2023**, *23*, 5632. [CrossRef]
13. Agrawal, D.K.; Usaha, W.; Pojprapai, S.; Wattanapan, P. Fall Risk Prediction Using Wireless Sensor Insoles with Machine Learning. *IEEE Access* **2023**, *11*, 23119–23126. [CrossRef]
14. Nadee, C.; Chamnongthai, K. An Ultrasonic-Based Sensor System for Elderly Fall Monitoring in a Smart Room. *J. Healthc. Eng.* **2022**, *2022*, 2212020. [CrossRef] [PubMed]
15. Zou, S.; Min, W.; Liu, L.; Wang, Q.; Zhou, X. Movement Tube Detection Network Integrating 3D CNN and Object Detection Framework to Detect Fall. *Electronics* **2021**, *10*, 898. [CrossRef]
16. Mei, X.; Zhou, X.; Xu, F.; Zhang, Z. Human Intrusion Detection in Static Hazardous Areas at Construction Sites: Deep Learning–Based Method. *J. Constr. Eng. Manag.* **2023**, *149*, 04022142. [CrossRef]
17. Delgado-Escano, R.; Castro, F.M.; Cozar, J.R.; Marin-Jimenez, M.J.; Guil, N.; Casilari, E. A crossdataset deep learning-based classifier for people fall detection and identification. *Comput. Methods Programs Biomed.* **2020**, *184*, 105265. [CrossRef] [PubMed]
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 27–29 October 2017; pp. 2961–2969.
19. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. *arxiv* **2016**, arXiv:1605.06409.
20. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
21. Krichen, M. Convolutional Neural Networks: A Survey. *Computers* **2023**, *12*, 151. [CrossRef]
22. Available online: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003393030-10/learning-modeling-technique-convolution-neural-networks-online-education-fahad-alahmari-arshi-naim-hamed-alqa (accessed on 5 March 2024).
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. 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.
25. Bochkovskiy, A.; Wang, C.Y.; Liao HY, M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
26. Zhang, L.; Ding, G.; Li, C.; Li, D. DCF-Yolov8: An Improved Algorithm for Aggregating Low-Level Features to Detect Agricultural Pests and Diseases. *Agronomy* **2023**, *13*, 2012. [CrossRef]
27. Lou, H.; Duan, X.; Guo, J.; Liu, H.; Gu, J.; Bi, L.; Chen, H. DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. *Electronics* **2023**, *12*, 2323. [CrossRef]
28. Guo, X.; Liu, S.; Li, L.; Qin, Q.; Li, T.; Guo, X.; Liu, S.; Li, L.; Qin, Q.; Li, T. Pedestrian detection algorithm in scenic spots based on improved YOLOv8. *Comput. Eng.* **2024**, 1–11. Available online: http://www.ecice06.com/CN/10.19678/j.issn.1000-3428.0068125 (accessed on 5 March 2024).
29. Cao, Y.; Xu, H.; Zhu, X.; Huang, X.; Chen, C.; Zhou, S.; Sheng, K. Improved Fighting Behavior Recognition Algorithm Based on YOLOv8: EFD-YOLO. *Comput. Eng. Sci.* **2024**, 1–14. Available online: http://kns.cnki.net/kcms/detail/43.1258.TP.20240126.0819.002.html (accessed on 5 March 2024).
30. Yang, Z.; Feng, H.; Ruan, Y.; Weng, X. Tea Tree Pest Detection Algorithm Based on Improved Yolov7-Tiny. *Agriculture* **2023**, *13*, 1031. [CrossRef]
31. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7369–7378.
32. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X.-Y. YOLO-FaceV2: A Scale and Occlusion Aware Face Detector. *arXiv* **2022**, arXiv:2208.02019.