applied sciences

MDPI

Article

PGDS-YOLOv8s: An Improved YOLOv8s Model for Object Detection in Fisheye Images

Degang Yang, Jie Zhou, Tingting Song *, Xin Zhang and Yingze Song

College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China; yangdg@cqnu.edu.cn (D.Y.); 2021110516034@stu.cqnu.edu.cn (J.Z.); 2021110516031@stu.cqnu.edu.cn (X.Z.); 2021110516019@stu.cqnu.edu.cn (Y.S.)

* Correspondence: ttsong@cqnu.edu.cn

Abstract: Recently, object detection has become a research hotspot in computer vision, which often detects regular images with small viewing angles. In order to obtain a field of view without blind spots, fisheye cameras, which have distortions and discontinuities, have come into use. The fisheye camera, which has a wide viewing angle, and an unmanned aerial vehicle equipped with a fisheye camera are used to obtain a field of view without blind spots. However, distorted and discontinuous objects appear in the captured fisheye images due to the unique viewing angle of fisheye cameras. It poses a significant challenge to some existing object detectors. To solve this problem, this paper proposes a PGDS-YOLOv8s model to solve the issue of detecting distorted and discontinuous objects in fisheye images. First, two novel downsampling modules are proposed. Among them, the Max Pooling and Ghost's Downsampling (MPGD) module effectively extracts the essential feature information of distorted and discontinuous objects. The Average Pooling and Ghost's Downsampling (APGD) module acquires rich global features and reduces the feature loss of distorted and discontinuous objects. In addition, the proposed C2fs module uses Squeeze-and-Excitation (SE) blocks to model the interdependence of the channels to acquire richer gradient flow information about the features. The C2fs module provides a better understanding of the contextual information in fisheye images. Subsequently, an SE block is added after the Spatial Pyramid Pooling Fast (SPPF), thus improving the model's ability to capture features of distorted, discontinuous objects. Moreover, the UAV-360 dataset is created for object detection in fisheye images. Finally, experiments show that the proposed PGDS-YOLOv8s model on the VOC-360 dataset improves mAP@0.5 by 19.8% and mAP@0.5:0.95 by 27.5% compared to the original YOLOv8s model. In addition, the improved model on the UAV-360 dataset achieves 89.0% for mAP@0.5 and 60.5% for mAP@0.5:0.95. Furthermore, on the MS-COCO 2017 dataset, the PGDS-YOLOv8s model improved AP by 1.4%, AP_{50} by 1.7%, and AP_{75} by 1.2% compared with the original YOLOv8s model.

Keywords: fisheye image; object detection; pooling; ghost module; Squeeze-and-Excitation (SE) block

Citation: Yang, D.; Zhou, J.; Song, T.; Zhang, X.; Song, Y. PGDS-YOLOv8s: An Improved YOLOv8s Model for Object Detection in Fisheye Images. *Appl. Sci.* **2024**, *14*, 44. https://doi.org/10.3390/app14010044

Academic Editor: João M. F. Rodrigues

Received: 18 November 2023
Revised: 15 December 2023
Accepted: 17 December 2023
Published: 20 December 2023

1. Introduction

Object detection is an important research direction in computer vision, and current object detectors usually detect conventional images with small viewpoints. With the continuous development of computer vision technology, the imaging range of captured images needs to be broader. Unmanned aerial vehicles (UAVs) at a higher altitude can access a richer range of information than ground-based shots [1,2]. The UAV expands the viewing angle by raising the height, and at the same time, the overall object within the image becomes smaller, which is difficult for current detectors to detect. Moreover, the fisheye lens, which has a viewing angle of 180° to 270°, has by far the largest viewing angle and can acquire rich visual information. It makes the fisheye camera have a wide range of applications, such as unmanned aerial vehicles [3–6], autonomous driving [7,8], robotics [9–11], and so on [12–15]. In this paper, instead of expanding the visual range

by elevating the height of the UAV, a UAV-mounted fisheye camera is used to obtain a 360° view without blind spots. The combination of a UAV and a fisheye camera makes the information acquisition more comprehensive, which is an important research area in the development of computer vision.

The combination of a UAV and a fisheye camera can obtain information about the field of view without blind spots. Therefore, a UAV equipped with a fisheye camera plays a vital role in object detection, environmental monitoring, agricultural inspection, urban planning, disaster prevention and relief, etc. For example, Barmpoutis et al. [3] used a 360-degree camera on a UAV to obtain a blind-free field of view, and this collocation can play a significant advantage in forest fire monitoring. The captured equirectangular images are first converted to stereoscopic images. Then, the DeepLab V3+ networks are utilized to segment the flames and smoke in the images, and a post-validation adaptive method is utilized to reduce the rate of false positives. This approach effectively achieves the detection and localization of fire regions and can play an active role in early fire detection. Furthermore, Luo et al. [4] used panoramic images captured by UAVs to address the cost and safety concerns associated with the structural health assessment of infrastructure. Multiple-projection methods are proposed to address the effects of panoramic image distortion. Then, deep neural networks are used to detect the damage to multiple steel surfaces in the 360° panoramic image. In addition, Gao et al. [5] used two fisheye cameras to obtain an omnidirectional visual range for aerial robots, which is beneficial for safe navigation in complex environments. Meanwhile, a dual-fisheye visual–inertial system (VINS) is proposed, which uses two fisheye cameras and an inertial measurement unit (IMU) to realize spherical omnidirectional sensing. The dual-fisheye omnidirectional visual–inertial state estimator is externally calibrated to optimize the fisheye image distortion problem. Yang et al. [6] constructed an autonomous landing system for UAVs that can land automatically in GPS-denied environments. The system combines a fisheye camera with a wide field of view and a stereo camera with depth imaging to obtain rich visual information, forming a hybrid camera array that is easier to pinpoint. In particular, the system employs YOLOv3 to directly detect objects in a fisheye image, robustly realizing autonomous landing on a moving unmanned ground vehicle (UGV).

Most existing object detectors in computer vision often detect regular images with small viewing angles. This paper investigates and uses a fisheye camera to further explore object detection for wide-view angle images. Fisheye cameras have a wide angle of view, and a UAV equipped with a fisheye camera can obtain a blind-spot-free view. However, distorted and discontinuous objects appear in the captured fisheye images due to the unique viewing angle of fisheye cameras. This poses a significant challenge to some existing object detectors. Therefore, in this work, based on the YOLOv8 model, which is one of the current state-of-the-art lightweight object detectors, this paper proposes a PGDS-YOLOv8s model to solve the problem of detecting distorted and discontinuous objects in fisheye images.

2. Related Works

2.1. Cameras for Object Detection

At present, object detection is a research hotspot in computer vision, and object detectors usually detect conventional life images taken with traditional perspective cameras. Perspective cameras are designed based on human vision and have a small viewing angle. Geiger et al. [16] presented the KITTI dataset, which captures 6 h of real-world traffic scenes using multiple sensor modalities such as high-resolution color and grayscale stereo cameras, a Velodyne 3D laser scanner, and a high-precision GPS/IMU inertial navigation system. The KITTI dataset provides images and object labels for automatic object detection research for driving. Mao et al. [17] created the ONCE dataset, which consists of one million LiDAR scenes and seven million corresponding camera images selected from 144 driving hours. This dataset is used for a 3D object detection task in a self-driving scenario. In addition, the images captured by UAVs contain rich point-of-view information. For example, Naude et al. [18] used a SkyReach BushCat light sport aircraft mounted with

Canon 6D digital single-lens reflex cameras to collect 2101 images. The Aerial Elephant dataset contains 15,511 bush elephants in their natural African habitat. A baseline algorithm for elephant detection is also trained and tested to demonstrate the feasibility of the task. Mou et al. [2] used a drone to collect 14,375 UAV aerial images. The WAID (Wildlife Aerial Images from Drone) dataset encompasses six wildlife species and multiple habitat types. This study brings new data to the field of UAV detection of wildlife. Meanwhile, the introduced SE-YOLO model effectively detects small objects in UAV images.

Object detectors often detect regular images with smaller viewing angles. However, people's requirements for images are not only intuitive and clear but also more concerned about the completeness and comprehensiveness of image information. The imaging range of the fisheye lens can reach 180°–270°, and the binocular fisheye camera can acquire a 360-degree field of view. Fisheye cameras effectively fulfill the need for a wide viewing angle. The Oxford RobotCar Dataset [19] contains nearly 20 million images collected from vehicle-mounted cameras and LIDAR, GPS, and INS ground truths. The cameras include mainly the Point Grey Grasshopper2 monocular camera and the Point Grey Bumblebee XB3 trinocular stereo camera with a 180° fisheye lens. The Oxford RobotCar Dataset is used for research on localization and mapping of self-driving vehicles. Yogamani et al. [20] presented the WoodScape dataset, which uses four fisheye cameras with 190° horizontal FOV, LiDAR, and other devices to acquire images. The WoodScape dataset provides labels for autonomous driving tasks, including semantic segmentation, monocular depth estimation, object detection, etc. However, the unique viewing angle of the fisheye camera causes problems such as significant image distortions, exaggerated relative sizes and distances of objects, and discontinuity of objects at the edges. Chiang et al. [21] proposed a pedestrian detection method for fisheye images in top view. The method generates multiple perspective views from the fisheye image and then detects the combined image of multiple perspective patches using existing detectors. Chen et al. [22] proposed a shallow Concatenated Feature Pyramid Network (CFPN). The proposed concatenated block further reduces the number of convolutional layers. Also, the concatenated approach effectively preserves the spatial information of smaller objects at the end of the network. CFPN is better at detecting small vehicles in fisheye cameras in real-time traffic flow. Arsenali et al. [23] proposed a multi-task network (MTL) to perform joint semantic segmentation, boundary prediction, and object detection on raw fisheye images. Two rotation-invariant object detection methods for fisheye images have also been explored, including YOLO-RotRect and YOLO-Circ. They effectively reduce the complexity of the network, but accurate estimation is still challenging. Wei et al. [24] proposed Rotation-Mask Deformable Convolution (RMDC) to address the problem of rotation and distortion in top-view fisheye images. The method adaptively rotates the convolution filter and introduces a center-fixed deformable convolution. The learning capability of the convolution kernel and the object detection accuracy of the top-view fisheye images are improved.

In addition, a UAV equipped with a fisheye camera can obtain a blind-free field of view. A UAV equipped with a fisheye camera plays an important role in environmental monitoring, disaster prevention and relief, object detection, etc. Barmpoutis et al. [3] utilized a 360-degree camera on a UAV to obtain a blind-free field of view, which can be used for a significant advantage in forest fire monitoring. The DeepLab V3+ network is utilized to segment flames and smoke in an image, which effectively achieves the detection and localization of fire areas. Yang et al. [6] constructed an autonomous landing system for UAVs. The system combines a fisheye camera with a wide field of view and a stereo camera with depth imaging to acquire rich visual information. In particular, the system employs YOLOv3 to directly detect objects in fisheye images, robustly realizing autonomous landing of mobile unmanned ground vehicles (UGVs).

Currently, there are few fisheye image datasets of real scenes. As the demand for a wide field of view increases, the fisheye image dataset of real scenes needs to be expanded. In addition, the wide viewing angle of fisheye cameras leads to distortion and discontinuity of objects in the captured fisheye images. Existing object detectors mainly detect regular

images, and distorted and discontinuous objects in fisheye images pose a challenge to object detectors.

2.2. Object Detection Methods

Deep-learning-based object detection frameworks can be categorized into two-stage object detection algorithms and single-stage object detection algorithms. The two-stage object detection algorithm consists of generating candidate regions and performing a classification process on them. Two-stage object detectors include R-CNN [25], Fast-RCNN [26], Mask R-CNN [27], etc. Girshick et al. [25] proposed the R-CNN. Object candidate regions are first generated on the image, then features are extracted using a convolutional neural network for the candidate regions, and a support vector machine classifier is applied for classification. Subsequently, Girshick et al. [26] proposed the Fast RCNN, which incorporates the ideas of SPP-net to improve R-CNN and introduces the pooling of regions of interest to unify the input size. Mask R-CNN [27] adds another parallel branch for pixel-level segmentation of object instances, which detects the objects and allows for objects to perform pixel-level segmentation. The two-stage object detection algorithm is more accurate. However, it is slower and still has obstacles in meeting the needs of complex object detection scenarios. The other is the single-stage object detection algorithm, which treats object detection as a one-time problem and quickly localizes and classifies objects end-to-end. Single-Shot MultiBox Detector (SSD) [28] is a classic single-stage object detection algorithm that guarantees detection speed while keeping the detection accuracy comparable to two-stage object detection algorithms. You Only Look Once (YOLO) [29] is an iconic algorithm for single-stage object detection that transforms the detection problem into a regression problem, but it is not highly effective in detecting small objects. YOLOv2 [30] uses Darknet-19 to extract object features and uses global average pooling and batch normalization to improve network convergence. YOLOv3 [31] employs a residual network module and uses Darknet-53 as a backbone to enhance object detection performance. YOLOv4 [32] uses Complete Intersection over Union (CIoU) loss for predictive frame filtering to improve the accuracy and robustness of the model. YOLOv5 [33] uses Feature Pyramid Network (FPN) and Pixel Aggregation Network (PANet) structures in the neck network, which is designed to be lightweight and faster. Li et al. [34] proposed YOLOv6, which injects a self-distillation strategy for classification and regression tasks. YOLOv7 [35] proposed E-ELAN, which does not change the original gradient path, uses group convolution to increase the cardinality of the added features, and combines different groups of features in a way that shuffles and merges the cardinality.

Although these object detectors perform well, considering the characteristics of distorted and discontinuous objects in fisheye images and the balance between speed and accuracy, this paper uses the YOLOv8s model for improvement and experimentation. YOLOv8 [36] is a fast, accurate, robust, and lightweight network built on YOLOv5. The main ideas of the YOLOv8 algorithm are as follows. First, the C2f module is designed, which refers to the ideas of the C3 module in YOLOv5 and Efficient Layer Aggregation Networks (ELANs). It makes the YOLOv8 model lightweight and effectively improves the object detection accuracy of the model. Next, YOLOv8 compares with YOLOv5, which removes the objectness branch and keeps the classification and regression branches. The decoupling structure in the detection section separates the regression and classification, and the Distributive Focal Loss (DFL) is used as the regression loss, allowing the network to quickly focus on the location distribution close to the object location. Then, YOLOv8s uses the Anchor-Free method instead of the Anchor-Base method to correct the object positions. The Anchor-Free method predicts the distance from the center to the bounding box after locating the object's center. In addition, the Task-Aligned Assigner selects positive samples based on the weighted scores of classification and regression [37]. YOLOv8 has five models with different scales. They are YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x [38]. The network architecture of YOLOv8 consists of Input, Backbone, Head, and Detect. The network model consists of these base modules. The CBS module con-

sists of Convolution (Conv2d), Batch Normalization (BatchNorm2d), and SiLU activation functions. UP is the up-sampling operation. The C2f module parallelizes more gradient flow branches to obtain more information. Spatial Pyramid Pooling Fast (SPPF) is a highly efficient pooling module for extracting and fusing features.

2.3. Current Issues in Object Detection for Fisheye Images

Currently, object detection in fisheye images still faces many challenges. For example, there are fewer publicly available fisheye image datasets in real scenes, and object distortions and discontinuities in fisheye images suffer from the detection performance of object detectors. In this paper, the following improvements are made to address these challenges:

1. The UAV-360 dataset is captured using a UAV-mounted fisheye camera, which contains 2045 equirectangular projection images converted from fisheye images. This dataset is beneficial to enhance the practical study of object detection in fisheye images.
2. Standard convolution makes it challenging to recognize the distorted and discontinuous objects in fisheye images, so two novel downsampling modules are proposed for the characteristics of fisheye images. The Max Pooling and Ghost's Downsampling (MPGD) module effectively extracts the essential feature information of distorted and discontinuous objects. Meanwhile, the Average Pooling and Ghost's Downsampling (APGD) module obtains rich global features and reduces the feature loss of distorted and discontinuous objects. In addition, compared to the convolutional layers in the downsampling stage of the original YOLOv8s, the two downsampling modules slightly reduce the parameters and computation of the model.
3. The complex background information in fisheye images degrades object detection performance in fisheye images. The proposed C2fs module uses the Squeeze-and-Excitation (SE) block to model the interdependencies between channels and to obtain richer information about the feature gradient flow. The C2fs module provides a better understanding of the contextual information in fisheye images. Meanwhile, the SPPFSE module adds an SE block after Spatial Pyramid Pooling Fast (SPPF). It uses global information to selectively emphasize basic features, thus improving the model's ability to capture distorted and discontinuous features.
4. The proposed PGDS-YOLOv8s model is a lightweight network structure that applies the above-proposed modules to the YOLOv8s model. The PGDS-YOLOv8s model effectively solves the problem of missed and wrong detection of distorted and discontinuous objects and obtains excellent object detection performance in fisheye images.

3. Materials and Methods

In this work, the PGDS-YOLOv8s model is proposed to solve the problem of detecting distorted and discontinuous objects in fisheye images. Figure 1 describes the steps involved in detecting fisheye images. First, the image and label files of the fisheye image dataset are prepared. Second, the proposed model is trained on the training and validation sets, and the weight files of the model are generated. Third, the proposed model is detected on the test set. Finally, the proposed model is evaluated and analyzed using the evaluation metrics.

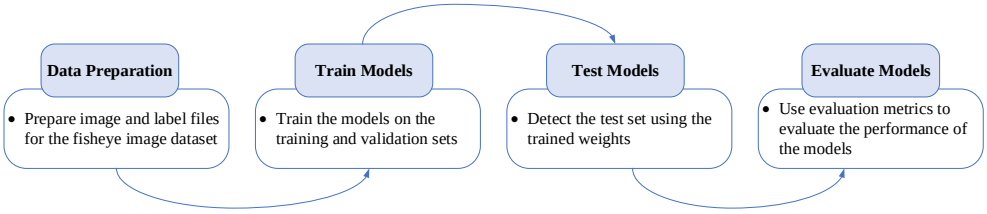


Figure 1. Detection process using the improved model.

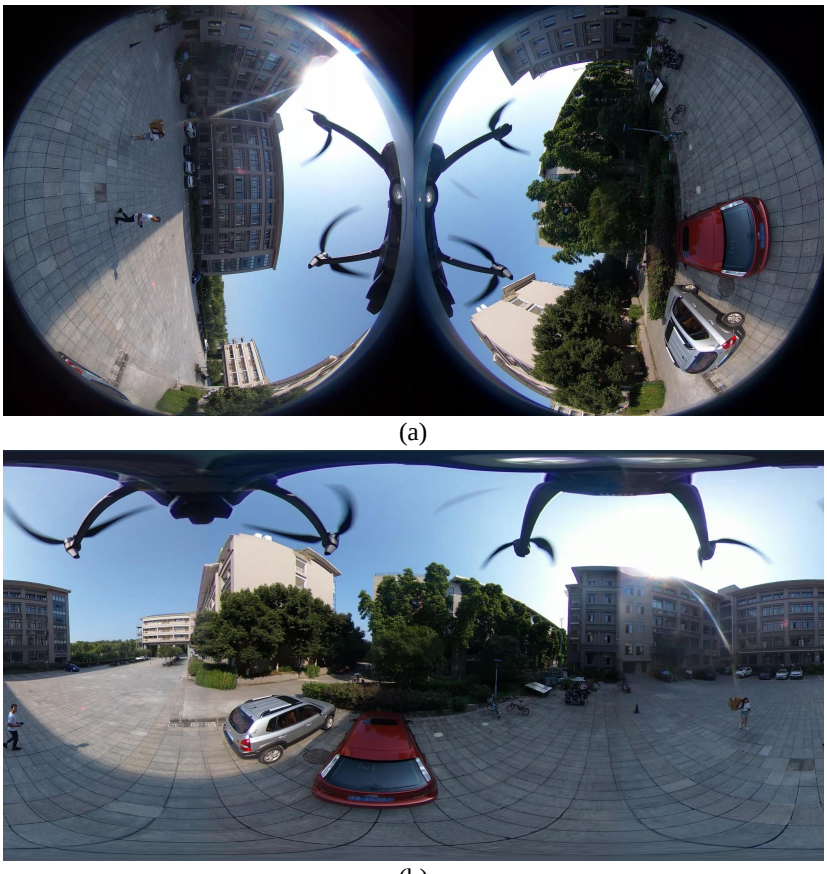
3.1. Data Preparation

This paper uses three datasets to evaluate the model performance, including the UAV-360, VOC-360, and MS-COCO 2017 datasets. Among them, the UAV-360 dataset contains

equirectangular projected images converted from fisheye images captured in real scenes, the VOC-360 dataset contains synthesized fisheye images, and the MS-COCO 2017 dataset contains regular images.

3.1.1. UAV-360 Dataset

In the data preparation stage, the DJI MAVIC 3 drone with the RICOH THETA V fisheye camera is used to collect fisheye images. The UAV-360 dataset contains equirectangular projected images converted from fisheye images captured with a UAV-mounted fisheye camera. The dataset has 2045 equirectangular projection images. Among them, the training set has 1430 images, the validation set has 410 images, and the test set has 205 images. In addition, the UAV-360 dataset has four categories, including people, cars, motorbikes, and buildings. The UAV-360 dataset is labeled with 12,530 objects, 3652 people, 4862 cars, 1063 motorbikes, and 2953 buildings. Figure 2a demonstrates the original binocular fisheye image. The two circular regions are the effective imaging regions. Figure 2b shows the equirectangular projection image converted from the binocular fisheye image. When a 360° image is rendered as a two-dimensional equirectangular projection image, it will inevitably have problems such as image distortion, relative size and distance of the objects being exaggerated, and incomplete objects at the edges.



(a)

(b)

Figure 2. Fisheye images. (**a**) Original binocular fisheye image; (**b**) equirectangular projection image converted from binocular fisheye image.

Figure 3 shows the labelimg tool drawing a rectangular bounding box for an object in an equirectangular projected image to obtain the XML file with the location information of the object. The XML file contains the image width and height, the object category, and the upper-left and lower-right coordinates of the rectangular bounding box of the object, etc. Finally, the center coordinates, width, and height of the object bounding box in the

XML file are normalized to the range [0, 1] and saved to the TXT file. The TXT file contains the category number, normalized center coordinates, normalized width, and normalized height. The TXT label file is used for object detection.

Figure 3. The labelimg tool labels objects in an equirectangular projected image, the green box is the bounding box for the labeling.

3.1.2. VOC-360 Dataset

The VOC-360 dataset is converted from the regular images of the VOC2012 dataset to fisheye images using an algorithm [39]. Figure 2a is the original binocular fisheye image. Figure 4 shows some synthesized fisheye images. The synthesized image is significantly similar to the actual fisheye image and suffers from severe distortion and discontinuous edge objects. The VOC-360 dataset has 20 categories, including people, motorcycles, cows, horses, birds, and so on. In this paper, 37,974 fisheye images are used as experimental data for object detection and randomly divided into training, validation, and test sets. Among them, the training set has 28,480 images, the validation set has 5696 images, and the test set has 3798 images.

(a) (b)

Figure 4. Some examples of the VOC-360 dataset. (**a**,**b**) are synthesized fisheye images.

3.1.3. MS-COCO 2017 Dataset

The MS-COCO 2017 dataset consists of a large regular image dataset containing 118k training images and 5k validation images. The eighty categories of this dataset include people, cars, motorcycles, birds, etc. It is commonly used to evaluate the performance of object detection models.

3.2. The Proposed Approach

Fisheye cameras have a wide viewing angle and can acquire rich perspective information. However, this leads to distorted and discontinuous objects in fisheye images. Therefore, this paper proposes a PGDS-YOLOv8s model based on the Novel Pooling and Ghost's Downsampling module and the SE module. The proposed PGDS-YOLOv8s model can effectively improve the detection performance of distorted and discontinuous objects in fisheye images. In addition, compared to the convolution of the downsampling stage of the original YOLOv8s, the two proposed Pooling and Ghost's Downsampling modules slightly reduce the parameters and computation of the model.

YOLOv8s, a state-of-the-art lightweight object detection network, is used as the base model in this paper. However, for the specificity of objects in fisheye images, the YOLOv8s model performs poorly in detection performance. In particular, the original YOLOv8s model employs a (3 × 3 convolutional kernel with a step size of 2) CBS module in the downsampling stage, which has difficulty in extracting features from distorted and discontinuous objects, thus losing a large amount of feature information. Therefore, two downsampling modules are proposed to solve the problem of feature loss for distorted and discontinuous objects. One of them is called the MPGD module, which is a downsampling module based on the Max Pooling and Ghost module, and the other is called the APGD module, which is a downsampling module based on the Average Pooling and Ghost module. The MPGD module efficiently extracts the key features of the distorted and discontinuous objects. The APGD module can obtain the global field of view, which reduces the feature loss of distorted and discontinuous objects. The two modules are used in the PGDS-YOLOv8s model to effectively improve the network's ability to recognize distorted and discontinuous objects in fisheye images. At the same time, compared to the CBS module of the original YOLOv8 model, the proposed MPGD and APGD modules reduce the parameters and computation of the model. In Backbone, the MPGD and APGD modules replace the (3 × 3 convolutional kernel with a step size of 2) CBS modules. The MPGD and APGD modules alternate for the downsampling operation, using the MPGD module first and then the APGD module. The APGD module improves the characteristic of the MPGD module that only extracts the most significant features, and it also focuses on the global features to maximize the retention of the information of the feature map. Meanwhile, in Head, the MPGD module replaces the two (3 × 3 convolution kernel with step size 2) CBS modules in the downsampling stage. Based on the features extracted from the backbone, the effective features in the fisheye image are further extracted. The MPGD and APGD modules have a large sensory field, which can effectively improve the performance of detecting distorted and discontinuous objects in the fisheye image.

In addition, the C2f module in the original YOLOv8 model uses a gradient shunt connection to obtain rich gradient flow information. To further enhance the model's understanding of the contextual information of fisheye images, the C2fs module is proposed. The C2fs module is the SE attention mechanism added to each Bottleneck of the C2f module. The C2fs module further enhances the ability to acquire channel features by retaining important feature information and suppressing unimportant feature information. Furthermore, the SPPFSE module is the addition of an SE layer after the Spatial Pyramid Pooling Fast (SPPF) module. It can use global information to selectively emphasize informative features, thereby improving the ability of the whole model to select and capture features. The proposed PGDS-YOLOv8s model is a lightweight network structure for fisheye images, which applies our proposed modules in the YOLOv8s model. Figure 5 shows the network structure of PGDS-YOLOv8s.

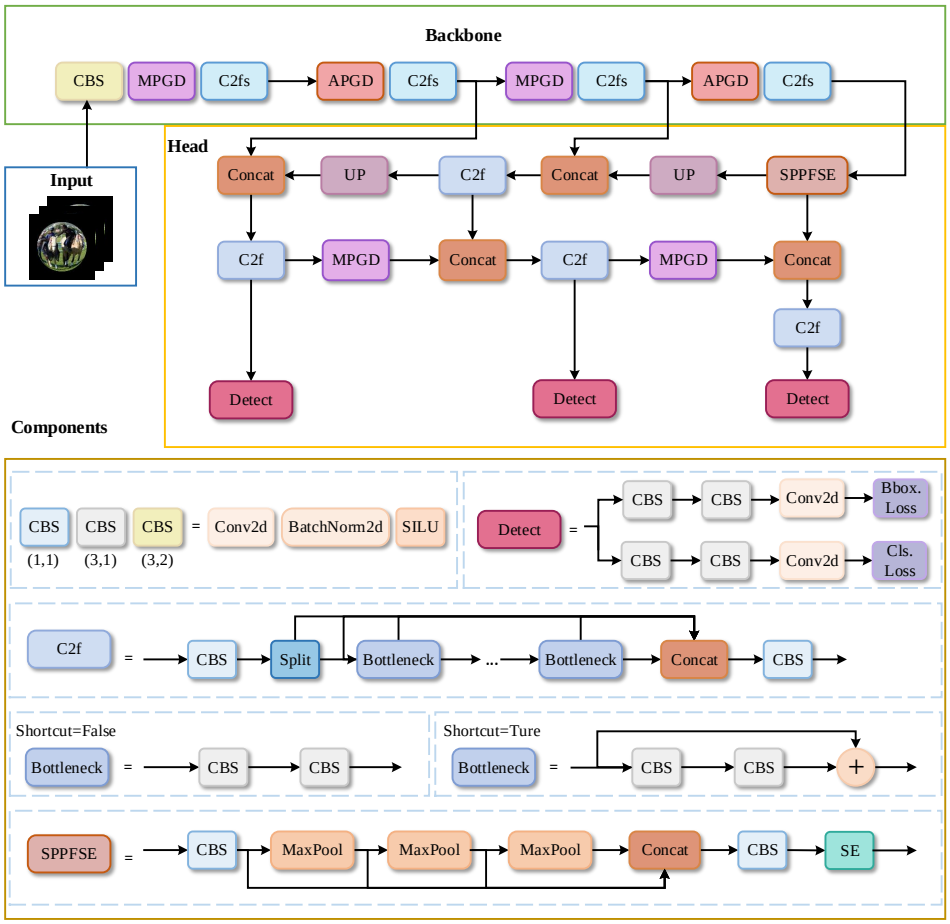


Figure 5. The network structure of PGDS-YOLOv8s.

3.2.1. Novel Downsampling Modules

Currently, the advanced lightweight YOLOv8s network is used to detect conventional images. However, when the YOLOv8s network detects fisheye images, the distorted and discontinuous features of objects in fisheye images make the results perform poorly. Conventional convolutional kernels have difficulty recognizing distorted and discontinuous objects, which severely loses the feature information of the objects. In particular, the original YOLOv8s model only uses one (3×3 convolutional kernels with a step size of 2) CBS to sample feature information in the downsampling stage. It is difficult to extract the features of distorted and discontinuous objects, losing many feature information of fisheye image objects. Therefore, two novel downsampling modules are proposed to solve the feature loss problem of distorted and discontinuous objects. Among them, the MPGD module is the Max Pooling and Ghost's Downsampling module, which can effectively extract the key features of distorted and discontinuous objects in fisheye images. The APGD module is the Average Pooling and Ghost's Downsampling module, which can effectively obtain the global features and reduce the feature loss of distorted and discontinuous objects.

Conventional convolution in the advanced lightweight YOLOv8s model has difficulty recognizing distorted and discontinuous objects in fisheye images. In particular, much feature information of distorted and discontinuous objects is lost in the downsampling stage. So, inspired by YOLOv7 [35] downsampling, pooling [40] is used to solve this problem. Pooling suppresses noise and reduces information redundancy. Max Pooling sparsifies the error, and Average Pooling equalizes the mistake. It effectively enhances the network's ability to extract distorted and discontinuous features in fisheye images. In addition, the pooling operation does not care about the specific location of the features and only abstracts the region's features. Features at the edges of the fisheye image are

less likely to be ignored, which is very beneficial for object detection in fisheye images. Meanwhile, pooling is often used to perform downsampling operations on the feature map, which can increase the network sensing field and improve the model's ability to perceive objects in fisheye images. In addition, the pooling layer has no parameters and does not require learning.

In addition, to further enhance the pooling module's understanding of contextual information and feature representations in fisheye images, the Ghost module [41] is added after the pooling module to extend the perceptual domain. This allows the model to efficiently acquire important information about the features of distorted and discontinuous objects. The Ghost module exploits the similarity of feature mappings and uses inexpensive linear operations to generate redundant feature mappings with key information. Compared to convolutional operations, the Ghost module effectively reduces the parameters and computational cost of the model, resulting in a more lightweight network model. First, the Ghost module generates half of the intrinsic feature maps using a 1 × 1 primary convolution. Then, for the intrinsic feature maps, cheap linear operations are performed on each channel to generate the other half of the ghost feature maps. The cheap linear operation uses a 5 × 5 convolution kernel to expand the receptive field. Finally, the two parts of the feature maps are concatenated to output the final feature maps. In addition, identity mapping is used to preserve the intrinsic feature maps. As shown in Figure 6, the regular convolutional layer and the Ghost module are illustrated to generate the same number of feature maps, respectively.

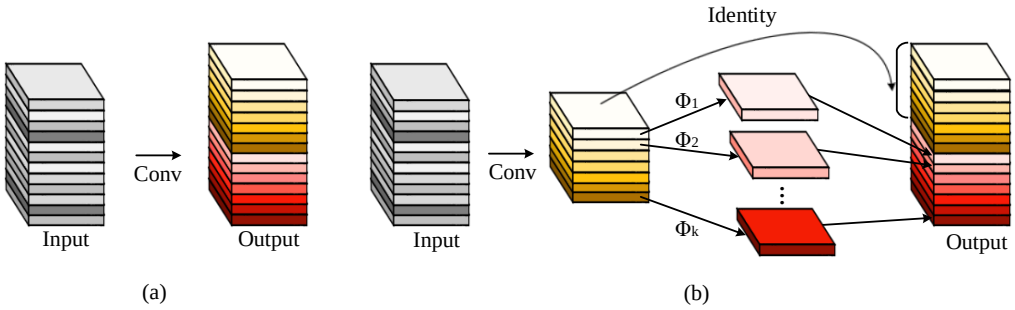


Figure 6. Illustration of the convolutional layer and Ghost module generating the same number of feature maps separately. (**a**) Convolutional layer; (**b**) Ghost module.

The proposed MPGD module uses two branches for the downsampling operation to effectively extract and retain the feature information of the feature map. Figure 7 illustrates the structure of the MPGD module. First, on one branch, the Max Pooling operation is used to extract the key feature information of the objects in the input feature map. Max Pooling does not care about the specific location of the features but only abstracts the critical features of the region, so it effectively preserves the key features of distorted and edge-discontinuous objects. At the same time, it sparsifies the error, suppresses noise, and reduces information redundancy. Subsequently, the Ghost module expands the receptive field to extract the essential features further and generate redundant features. The Ghost module generates intrinsic features using a (1 × 1 convolutional kernel with a step size of 1) CBS module. Then, a (5 × 5 convolutional kernel with a step size of 1) CBS module is used to expand the receptive field. Second, to further preserve the original feature information of the fisheye image, another branch uses a (1 × 1 convolution kernel with a step size of 1) CBS module to extract comprehensive feature information, and then, a (3 × 3 convolution kernel with a step size of 2) CBS is used for the downsampling operation. Finally, the branch information with local key features and global features is concatenated. The MPGD module acquires rich essential feature information during the downsampling process, and it maximizes the retention of critical features in the object in the fisheye image.

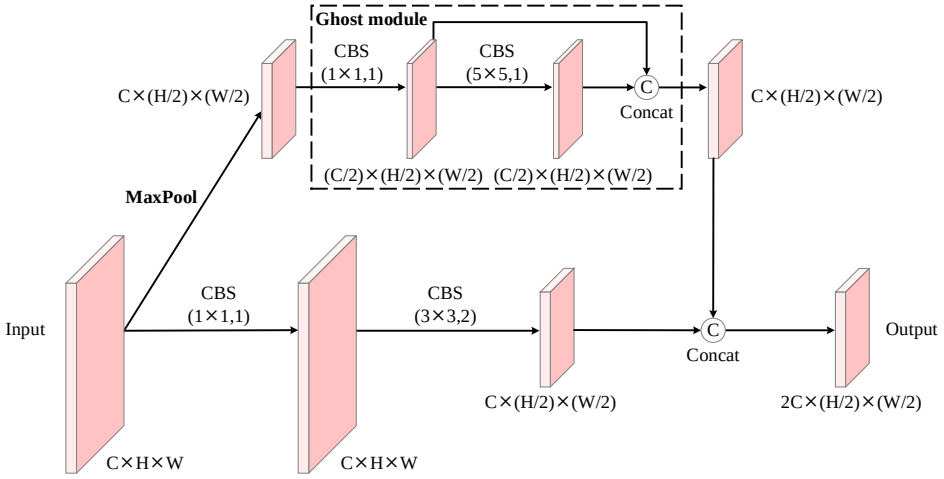


Figure 7. The structure of the MPGD module.

In addition, the APGD module is proposed based on the MPGD module to obtain more comprehensive fisheye image feature information. As shown in Figure 8, the APGD module is to replace the Max Pooling of the MPGD module with Average Pooling, and the other structures are the same. The Average Pooling can equalize the error in the downsampling process so that the APGD module can obtain rich global features. The MPGD module can effectively extract the key features in the fisheye image, and the APGD module can obtain rich global features in the fisheye image. The MPGD and APGD modules are used alternately in Backbone, which can effectively improve the overall feature extraction ability of the improved model for fisheye images.

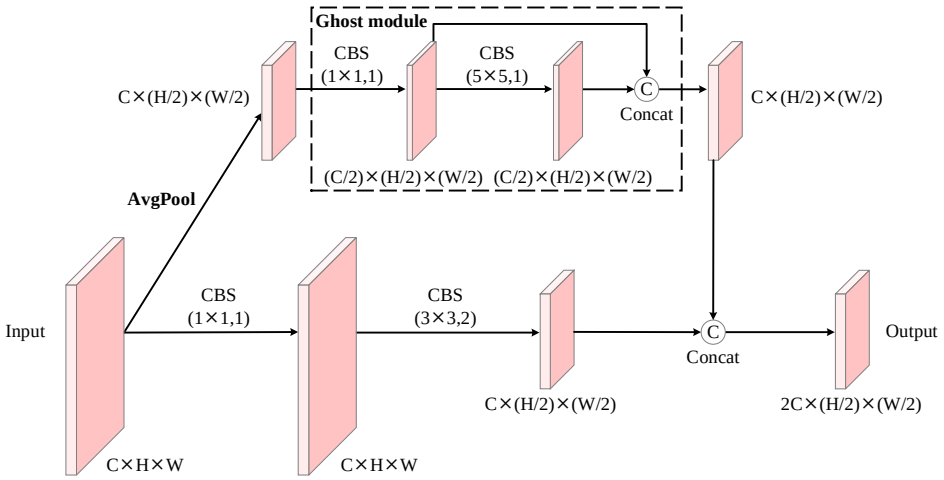


Figure 8. The structure of the APGD module.

As shown in Figures 7 and 8, given the input data $X \in \mathbb{R}^{C \times H \times W}$, C denotes the number of input channels, and H and W denote the height and width of the input data, respectively. First, MaxPool and AvgPool are uniformly denoted as Pool. Figure 7 adopts (pooling kernel size of 2 × 2) MaxPool, and Figure 8 adopts (pooling kernel size of 2 × 2) AvgPool. $Y \in \mathbb{R}^{C \times (H/2) \times (W/2)}$ denotes the feature map of the output of the pooling operation:

$$Y = Pool(X) \tag{1}$$

Subsequently, the pooled output feature map Y is taken as input; m intrinsic feature maps Y' are generated by primary convolution [41]; and $Y' \in \mathbb{R}^{H' \times W' \times m}$, where $*$ is the

convolution operation, $f \in \mathbb{R}^{C\times k\times k\times m}$ is the filters in the convolution layer, $k \times k$ indicates the size of the filter f, and the bias terms are omitted. The equation is expressed as follows:

$$Y' = Y * f \tag{2}$$

The s ghost feature maps Y'' are obtained for each intrinsic feature in Y' using cheap linear operations, where y'_i is the i-th intrinsic feature map in Y', $\Phi_{i,j}$ is the cheap linear operation, y_{ij} is the j-th ghost feature map generated by the i-th intrinsic feature map, and $\Phi_{i,s}$ is the constant mapping used to preserve the intrinsic feature maps. The linear operation Φ is performed for each channel, and its computational cost is much lower than that of ordinary convolution. The equation is expressed as follows:

$$y_{ij} = \Phi_{i,j}(y'_i), \quad \forall i = 1,\ldots,m, \quad j = 1,\ldots,s \tag{3}$$

Then, the intrinsic feature maps Y' and the ghost feature maps Y'' are concatenated to obtain the feature map $Y1$. The equation is expressed as follows:

$$Y1 = Concat(Y', Y'') \tag{4}$$

In another branch, $f' \in \mathbb{R}^{C\times k\times k\times n}$ is the filters in the convolution layer, n denotes the number of input feature maps, $k \times k$ indicates the size of the filter f', and the input feature map X is first processed using a 1 × 1 filter f' to retain comprehensive feature information. Then, a 3 × 3 filter f' performs a downsampling operation to obtain the feature map $Y2$.

$$Y2 = X * f' * f' \tag{5}$$

Finally, our proposed downsampling module is to concatenate the feature maps $Y1$ and $Y2$ on the two branches to obtain the feature map $Y3$ with rich feature information.

$$Y3 = Concat(Y1, Y2) \tag{6}$$

3.2.2. C2fs Module

The exaggerated relative size and distance of objects and complex background information in fisheye images lead to further degradation of object detection performance. To further optimize object detection performance in fisheye images, the Squeeze-and-Excitation (SE) block is introduced in this paper [42]. The SE block selectively emphasizes the information features using the global information, which will pay more attention to the information channel features and effectively improve the detection performance of the objects in fisheye images.

As shown in Figure 9, the SE block contains Squeeze-and-Excitation operations. First, the global average pooling in the squeeze operation processes the spatially informative features of the feature map into channel informative features to obtain the global features on the channels. This is followed by the excitation operation, which learns the nonlinear relationships between channels and captures the interdependencies between channels to recalibrate the channel features adaptively. Finally, the generated scalar is multiplied by the original feature map for feature fusion to obtain the final features. The SE block uses global information to emphasize informative features in the fisheye image selectively. It improves sensitivity to features in the fisheye image and suppresses less useful features.

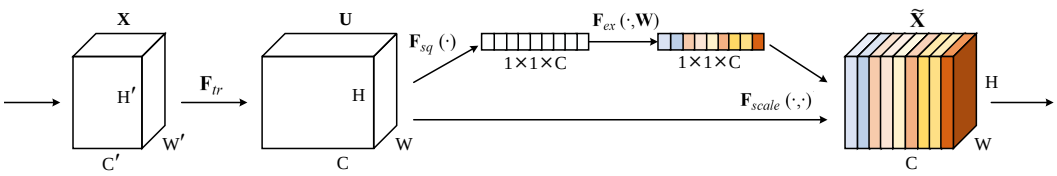


Figure 9. A Squeeze-and-Excitation block.

The $\mathbf{F}_{tr}$ operation generates a feature map $\mathbf{U}$ from the input feature map $\mathbf{X}$ with a convolution operation, where $\mathbf{X} \in \mathbb{R}^{H' \times W' \times C'}$ and $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$. $\mathbf{F}_{sq}$ denotes the squeeze operation. The squeeze performs the global average pooling operation on the feature map to generate the channel-wise statistics $\mathbf{z}$ [42], where $\mathbf{z} \in \mathbb{R}^{C}$, which represents the global perceptual field of each channel using a numerical value, where the equation for the c-th element of $\mathbf{z}$ is expressed as follows:

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{7}$$

$\mathbf{F}_{ex}$ denotes the excitation operation, and the excitation uses two fully connected layers to parameterize the gating mechanism and to capture channel-wise dependencies. Finally, the scalar $\mathbf{s}$ is obtained. δ denotes the ReLU function, and σ denotes the Sigmoid function. $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. r denotes the reduction ratio. The equation is expressed as follows:

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \tag{8}$$

The $\mathbf{F}_{scale}$ operation multiplies the scalar s_c with the corresponding channel of the feature map $\mathbf{u}_c$ to obtain the final feature map $\widetilde{\mathbf{x}}$, with $\mathbf{u}_c \in \mathbb{R}^{H \times W}$. The equation is expressed as follows:

$$\widetilde{\mathbf{x}}_c = \mathbf{F}_{scale}(\mathbf{u}_c, s_c) = s_c \cdot \mathbf{u}_c \tag{9}$$

The C2f module in the original YOLOv8 model uses a gradient-splitting connection to obtain rich gradient flow information. To further enhance the model's understanding of the contextual information of fisheye images, the C2fs module is proposed. As shown in Figure 10, the C2fs module replaces the Bottleneck of the C2f module with the SE-Bottleneck. The SE-Bottleneck adds the SE attention mechanism to the Bottleneck. One branch first extracts the effective features using two 3 × 3 convolutional layers and then adaptively recalibrates the features using the SE module modeling the interdependence of the channels. The other branch is the original features. Finally, the features from these two branches are fused. The SE-Bottleneck can acquire channel features efficiently and learn deeper features in fisheye images. The C2fs module employs an SE-Bottleneck, which allows the C2fs module to efficiently fuse different levels of features and extract richer contextual information from fisheye images. As shown in Figure 5, all the C2f modules are replaced with C2fs modules in Backbone to further enhance the performance of the improved model for object detection against fisheye images. Meanwhile, the SPPFSE module in Head is an SE block added after the Spatial Pyramid Pooling Fast (SPPF) module. It uses global information to selectively emphasize informative features based on the fusion of deep and shallow information by SPPF, thus improving the ability of the whole model to select and capture features.

3.3. Evaluation Metrics

In this study, Precision, Recall, mAP@0.5, mAP@0.5:0.95, GFLOPs, and FPS are used to evaluate object detection performance for fisheye images in the VOC-360 and UAV-360 datasets. Among them, mAP@0.5 denotes the Mean Average Precision across all categories at an IoU threshold of 0.5. mAP@0.5:0.95 indicates the Mean Average Precision for IoU thresholds from 0.5 to 0.95 in steps of 0.05. FPS is the number of image frames processed per second. In addition, AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L are used to evaluate object detection performance for regular images in the MS-COCO 2017 dataset. Performance is evaluated using three IoU threshold types of AP, including AP (IoU threshold average), AP_{50} (IoU threshold = 0.50), and AP_{75} (IoU threshold = 0.75). At the same time, three criteria are used to evaluate the accuracy of small, medium, and large objects corresponding to AP_S, AP_M, and AP_L [43].

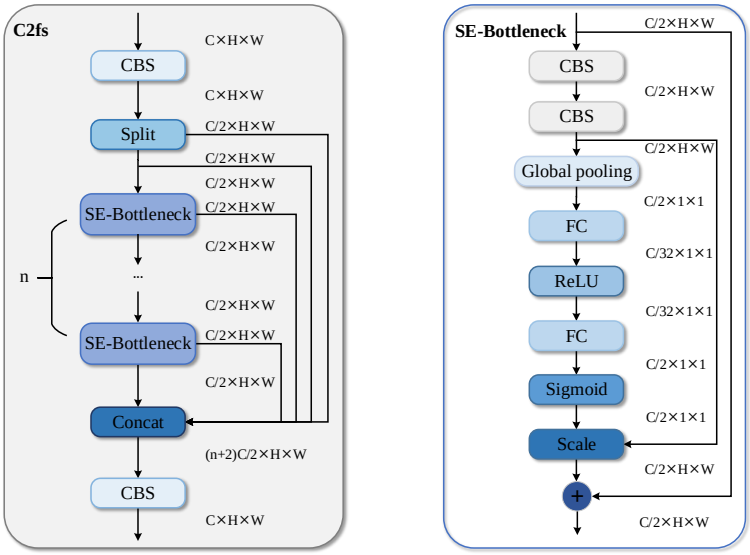


Figure 10. The structure of the C2fs module and the SE-Bottleneck module.

4. Results

4.1. Implementation Details

The RICOH THETA V Binocular Fisheye Camera has a 360° field of view and is compact and portable. It can stabilize the image even in motion, retaining more detailed information and presenting a more realistic visual effect. At the same time, the fisheye camera is connected to the cell phone via wireless LAN to realize real-time, high-speed image transmission. Fisheye images captured in real time are displayed on the cell phone. Therefore, this paper selects the RICOH THETA V binocular fisheye camera to capture fisheye images with a 360° view angle. In addition, the DJI MAVIC 3 drone is easy and safe to operate. Meanwhile, it has a range of 46 min. Moreover, it can detect objects in all directions during flight and dexterously bypass obstacles for advanced intelligent return. So, the DJI MAVIC 3 drone with the RICOH THETA V fisheye camera is used to collect data. The mounting method is used to fix the fisheye camera to the bottom of the drone using brackets.

In addition, all experiments are performed in Pytorch 1.10.0 and cuda 11.3 environment. A NVIDIA GeForce RTX 3080 Ti GPU device is used to train the VOC-360 and UAV-360 datasets, and the NVIDIA GeForce RTX 3090 GPU device is used to train the MS-COCO 2017 dataset.

The hyperparameters on the three datasets are set as follows. Firstly, all models in the VOC-360 dataset are trained with 300 epochs, the batch size is set to 8, and the initial learning rate is set to 0.005. Secondly, all models in the UAV-360 dataset are trained with 100 epochs, the batch size is set to 8, and the initial learning rate is set to 0.02. In addition, all models are trained for 100 epochs on the MS-COCO 2017 dataset, the batch size is set to 32, and the initial learning rate is set to 0.01.

4.2. Ablation Experiment

The experimental parameters and results of the improved models are compared with the original models on the VOC-360 dataset. The experimental results in Table 1 demonstrate that the improved models significantly improve object detection performance for fisheye images.

Table 1. Comparison of improved model results on the VOC-360 dataset.

Methods	Lr	Params (M)	GFLOPs	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FPS
YOLOv8s	0.01	11.14	28.7	73.0	52.6	59.4	35.3	286
YOLOv8s	0.005	11.14	28.7	75.2	58.0	63.8 (+4.4)	40.2 (+4.9)	286
YOLOv8s + MPGD + APGD	0.005	10.79	28.5	82.6	72.0	78.7 (+19.3)	62.4 (+27.1)	256
YOLOv8s + C2fs + SPPFSE	0.005	11.14	28.7	80.8	65.5	72.8 (+13.4)	51.4 (+16.1)	286
PGDS-YOLOv8s	0.005	10.79	28.5	85.1	71.3	79.2 (+19.8)	62.8 (+27.5)	250

4.2.1. Comparison of Improved Model Results on the VOC-360 Dataset

The VOC-360 dataset contains synthetic fisheye images created by sampling patches in regular images and then using an algorithm to simulate the spherical viewing angle of the original fisheye images. Therefore, most synthetic fisheye images have the characteristics of the original fisheye image's viewing angle, object distortion, objects near large and far small, edges produce discontinuous objects, etc. The YOLOv8s model makes recognizing objects in the regular image easy. For special characteristics of the fisheye image, the YOLOv8s model does not easily recognize distorted and discontinuous objects in the fisheye image, resulting in a lower overall object detection accuracy. At the same time, when the learning rate is 0.01, the learning rate is large, leading to more misdetections and omissions.

When the YOLOv8s model is trained on the VOC-360 dataset, the default initial learning rate significantly impacts detection accuracy. Larger learning rates converge quickly at first but may fail to converge to the optimal value. Therefore, the initial learning rate of all the improved models is set to 0.005. The appropriate initial learning rate allows all the improved models to optimize the original model's false and missed detection rates.

The MPGD and APGD modules can effectively extract distorted and discontinuous features of the fisheye image in the downsampling stage. At the same time, compared to the convolutional layers in the downsampling stage of the original YOLOv8s, the two proposed downsampling modules slightly reduce the parameters and computation of the model. YOLOv8s model introduces the MPGD and APGD modules, which improves the mAP@0.5 by 19.3% and mAP@0.5:0.95 by 27.1% compared with the original model. In addition, the SE module models the interdependence of the channels and adaptively recalibrates the features. This allows the C2fs module to acquire richer contextual information, and the SPPFSE module improves the ability of the whole model to capture distorted and discontinuous object features. The YOLOv8s model introduces the C2fs and SPPFSE modules, which improves the mAP@0.5 by 13.4% and the mAP@0.5:0.95 by 16.1% compared to the original model. The PGDS-YOLOv8s model combines the advantages of several modules and obtains the best results in object detection in fisheye images. Compared with the original model, the PGDS-YOLOv8s model improves mAP@0.5 by 19.8% and mAP@0.5:0.95 by 27.5%.

4.2.2. Performance Comparison of Improved Models on the VOC-360 Dataset

Figure 11 shows the test results of some fisheye images in the VOC-360 test set. The advantages of the MPGD and APGD modules are demonstrated in Figure 11b. The MPGD module is good at extracting effective features in fisheye images, while the APGD module better solves the problems of missed and wrong detection. Introducing the MPGD and APGD modules into the YOLOv8s model significantly improves the detection accuracy of distorted and discontinuous objects in fisheye images. At the same time, the leakage detection is effectively improved, and the detection frame region is more accurate. In addition, as shown in Figure 11c, the YOLOv8s model, with the introduction of the C2fs and SPPFSE modules, which extracts informative channel features and focuses on regions of interest, effectively improves the detection accuracy of objects in fisheye images. Finally, as shown in Figure 11d, the PGDS-YOLOv8s module introduces the proposed modules, which perform very well in the overall performance. Its detection frames are more accurate and detailed, resulting in excellent recognition of objects of different scales in fisheye images.

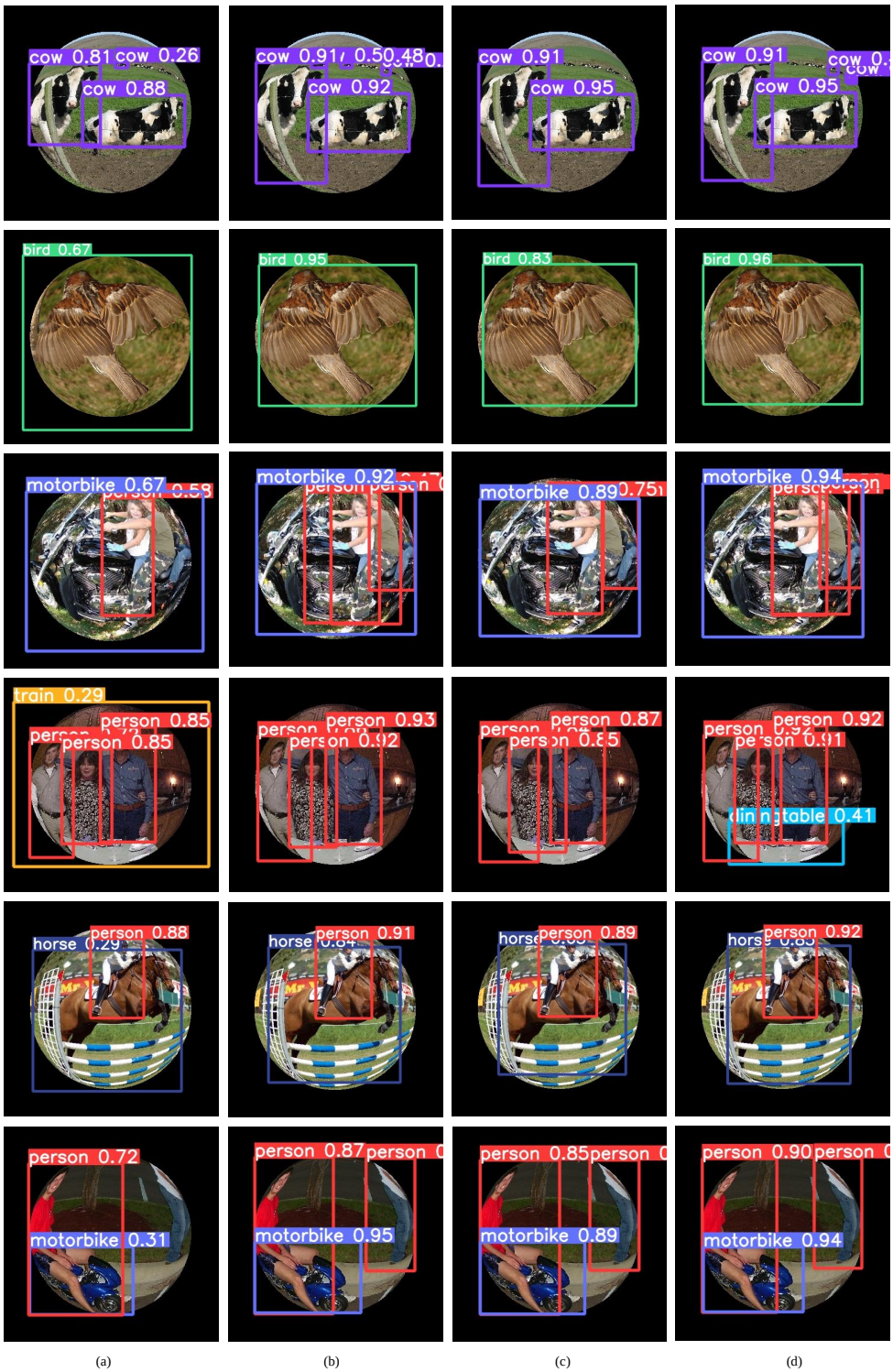


Figure 11. Comparison of test results of four models on the VOC-360 test set. (**a**) YOLOv8s model; (**b**) YOLOv8s + MPGD + APGD model; (**c**) YOLOv8s + C2fs + SPPFSE model; (**d**) PGDS-YOLOv8s model.

4.3. Comparison with Several Advanced Models on the VOC-360 Dataset

Simultaneously, object detection is performed with the YOLOv3-tiny, YOLOv5s, YOLOX-s, YOLOX-m, YOLOv6-S, YOLOv6-M, YOLOv7-tiny, YOLOv7, and YOLOv8m models on the VOC-360 dataset. The detection results are compared with those of the YOLOv8s and PGDS-YOLOv8s models. The comparison results on the VOC-360 dataset

are presented in Table 2. In particular, the YOLOv8s model is more challenging to adapt to object detection of distorted images and shows lower detection accuracy on the VOC-360 dataset. Compared with the YOLOv8s model, the proposed PGDS-YOLOv8s model has a reduced number of parameters and computation and significantly improves object detection accuracy for fisheye images. The PGDS-YOLOv8s model outperforms other models such as YOLOv3-tiny, YOLOv5s, YOLOX-s, YOLOX-m, YOLOv6-S, YOLOv6-M, YOLOv7-tiny, and YOLOv8m at mAP@0.5 and mAP@0.5:0.95. The PGDS-YOLOv8s model for mAP@0.5 is lower than the YOLOv7 model. However, the PGDS-YOLOv8s model for mAP@0.5:0.95 is better than the YOLOv7 model. Moreover, the YOLOv7-tiny model performs best on FPS. The experimental results show that the PGDS-YOLOv8s model outperforms most other models for object detection in fisheye images.

Table 2. Comparison with several advanced models on the VOC-360 dataset.

Methods	Size	Params (M)	GFLOPs	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FPS
YOLOv8s	640	11.14	28.7	59.4	35.3	286
YOLOv3-tiny	640	8.71	13.1	57.7	27.2	476
YOLOv5s	640	7.07	16.1	63.9	34.3	357
YOLOX-s	640	8.95	26.8	73.2	52.4	193
YOLOX-m	640	25.29	73.79	78.0	60.2	91
YOLOv6-S	640	18.51	45.2	77.4	58.9	154
YOLOv6-M	640	34.82	85.67	78.9	61.1	107
YOLOv7-tiny	640	6.06	13.3	67.9	45.4	588
YOLOv7	640	37.29	105.4	80.6	61.9	182
YOLOv8m	640	25.87	79.1	66.6	42.3	145
PGDS-YOLOv8s	640	10.79	28.5	79.2	62.8	250

4.4. Comparison with Several Advanced Methods on the UAV-360 Dataset

Table 3 shows a comparison of the experimental results of the proposed method with the state-of-the-art object detection methods for fisheye images. Among them, Kim et al. [44] proposed three multi-scale feature connection models for detecting fisheye images, including Long-Skip Concatenation Model (LCat), Short-Skip Concatenation Model (SCat), and Short–Long-Skip Concatenation Model (SLCat). In addition, the edge continuity distortion-aware block (ECDAB) [45] and SphereConv [46] are introduced into the YOLOv8s model for comparison experiments with the proposed model, respectively. ECDAB mitigates object discontinuities and distortions in fisheye images by recombining and segmenting features. In addition, SphereConv resolves the object edge discontinuities and distortions in fisheye images by resampling. Our proposed method employs MPGD and APGD modules in the downsampling stage to expand the receptive field, while the C2fs module is utilized to acquire rich feature gradient flow information further. The features of distorted and discontinuous objects in fisheye images are effectively preserved to obtain good detection results. The experimental results demonstrate that all the object detection methods for fisheye images on the UAV-360 dataset have good detection accuracy, and the detection performance of our proposed method outperforms all the others. Among them, the FPS of the YOLOv8s model with the introduction of the MPGD and APGD modules reaches 208. The PGDS-YOLOv8s model achieves 89.0% for mAP@0.5 and 59.9% for mAP@0.5:0.95. In particular, the YOLOv8s model introducing the C2fs and SPPFSE modules achieves 60.5% for mAP@0.5:0.95. With the complex background in the equirectangular projected image, the C2fs and SPPFSE modules allow the model to focus more on the informative channel features, further enhancing the model's understanding of the contextual information in the equirectangular projected image. The method obtains the best accuracy for object detection on the UAV-360 dataset.

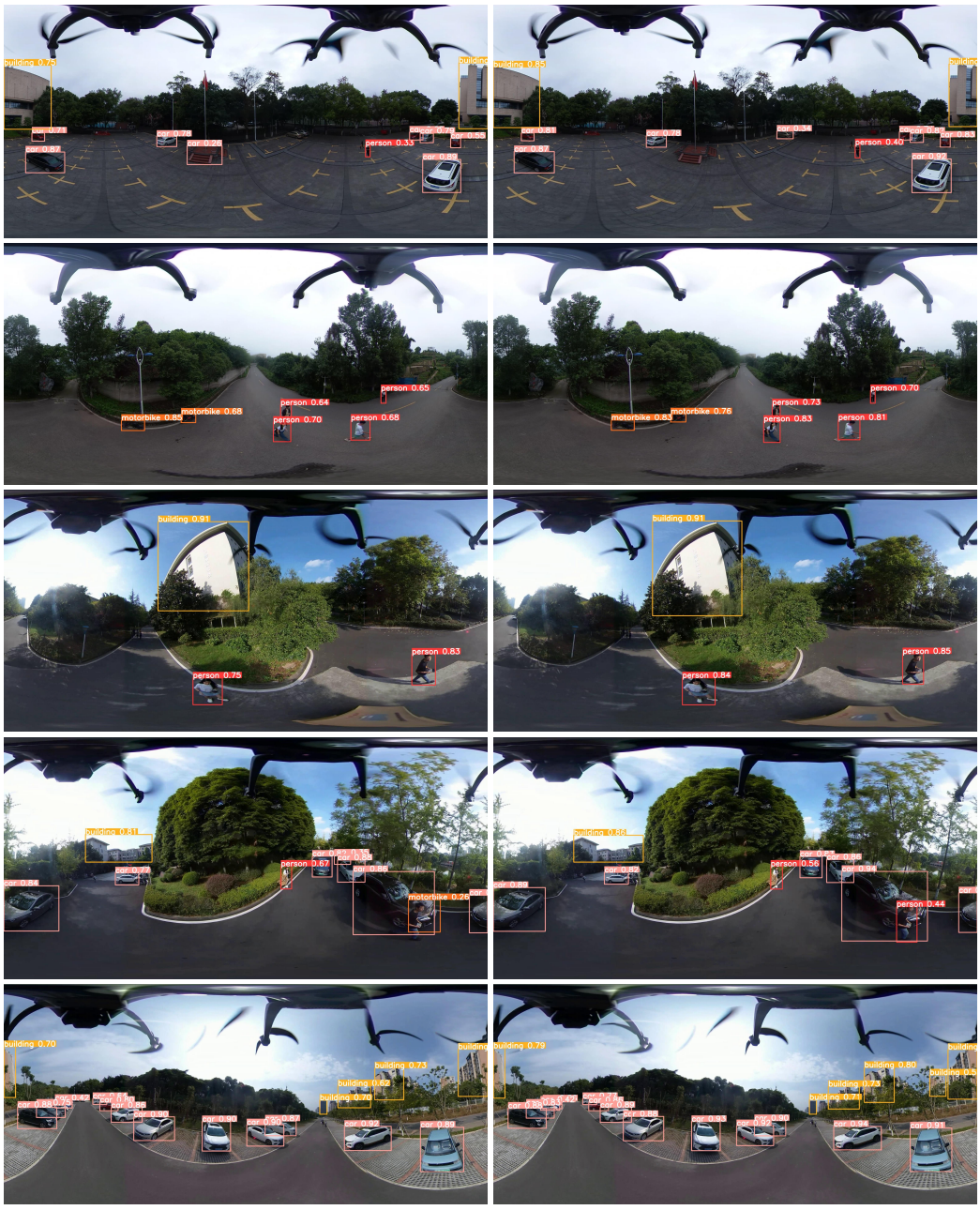
Table 3. Comparison with several advanced methods on the UAV-360 dataset.

Methods	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FPS
YOLOv8s	86.6	79.9	87.1	57.7	189
LCat	83.2	78.6	84.6	48.9	127
SCat	84.9	82.1	87.0	55.2	119
SLCat	82.4	78.8	84.9	48.7	123
YOLOv8s + ECDAB	85.0	83.0	87.8	58.1	185
YOLOv8s + SphereConv	84.9	82.2	87.6	57.9	175
YOLOv8s + MPGD+APGD	87.4	81.1	88.7	60.2	208
YOLOv8s + C2fs + SPPFSE	85.5	83.7	89.0	60.5	169
PGDS-YOLOv8s	86.4	82.9	89.0	59.9	179

Figure 12 shows the test results for some equirectangular projection images in the UAV-360 test set. Images captured by the fisheye camera show that objects closer to the lens are larger and significantly distorted, while objects farther away from the lens are smaller. In the equirectangular projection images, objects at the poles of the image are particularly distorted, and objects located at the left and right edges of the image are discontinuous. The original YOLOv8 model does not perform well in dealing with these problems. For example, in the first scene of the first row, the detection accuracy of discontinuous buildings on the left and right sides of the equirectangular projection image is low. In the first scene of the first row, the flagpole platform is incorrectly detected as a car, and some of the vehicles cannot be recognized. In the first scene of the fourth row, a person is incorrectly detected as a motorbike. Compared with the original YOLOv8s model, the PGDS-YOLOv8s model can accurately recognize more objects and has higher object detection accuracy. The MPGD module can efficiently extract the key feature information of the distorted and discontinuous objects in the fisheye image. The APGD module can obtain the global field of view and reduce the feature loss of distorted and discontinuous objects. The two downsampling modules prevent the features at the edges of the fisheye image from being easily ignored, which is very favorable for object detection in fisheye images. The C2fs module effectively fuses different levels of features and extracts richer contextual information in fisheye images. It can focus on the region of interest and effectively improve object detection accuracy. The proposed PGDS-YOLOv8s model performs well in object detection in fisheye images.

4.5. Improved Models' Comparison on the MS-COCO 2017 Dataset

Our proposed method demonstrates excellent performance in object detection for fisheye images. To further validate the broad applicability of the proposed method, the MS-COCO 2017 dataset is used for experiments. Compared with the original YOLOv8s model, the improved models exhibit more excellent object detection performance. Among them, the YOLOv8s model introducing MPGD and APGD modules reduces the feature loss in the downsampling stage and performs better in detecting smaller objects. The YOLOv8s model introducing the C2fs and SPPFSE modules pays more attention to the informative channel features and performs better on larger objects. The PGDS-YOLOv8s model combines the advantages of the proposed modules to show the best detection performance. Table 4 shows that several evaluation performance metrics of the PGDS-YOLOv8s model on the MS-COCO 2017 dataset perform very well. Among them, AP improved by 1.4%, AP_{50} improved by 1.7%, AP_{75} improved by 1.2%, AP_S improved by 2.5%, AP_M improved by 1.1%, and AP_L improved by 1.7%. The experiments demonstrate that the proposed model still enhances object detection for images on the MS-COCO 2017 dataset.



(**a**) (**b**)

Figure 12. Comparison of the test results of the two models on the UAV-360 test set. (**a**) YOLOv8s model; (**b**) PGDS-YOLOv8s model.

Table 4. Improved models comparison on the MS-COCO 2017 dataset.

Methods	Params (M)	GFLOPs	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)	FPS
YOLOv8s	11.16	28.7	42.1	58.4	45.9	22.9	46.7	58.0	303
YOLOv8s + MPGD + APGD	10.81	28.5	43.1	59.4	47.0	24.3	47.7	58.5	286
YOLOv8s + C2fs + SPPFSE	11.17	28.7	42.4	58.9	46.3	23.2	47.2	58.5	303
PGDS-YOLOv8s	10.81	28.5	43.5	60.1	47.1	25.4	47.8	59.7	294

5. Discussion

Most existing object detectors often detect regular images with small viewing angles. To further explore the object detection of wide-view angle images, this paper uses images with wide-view angles captured by a fisheye camera mounted on a UAV. However, the

unique viewing angle of the fisheye camera leads to problems such as object distortion and discontinuity of objects at the edges in fisheye images. This poses a significant challenge to advanced object detectors. Therefore, to solve these problems, this paper proposes a PSG-YOLOv8s model for object detection in fisheye images. The MPGD module effectively extracts key feature information of distorted and discontinuous objects. The APGD module acquires rich global features and reduces feature loss of distorted and discontinuous objects. In addition, the C2fs module models the interdependence of the channels to acquire richer gradient flow information of the features. Subsequently, the SPPFSE module further improves the model's ability to capture features of distorted and discontinuous objects. The experimental results in this paper demonstrate that the proposed method can extract distorted and discontinuous features effectively, thus obtaining excellent detection performance. The proposed model is compared with some current state-of-the-art object detectors on the VOC-360 dataset. For fisheye images with a spherical viewing angle, the accuracy of the PGDS-YOLOv8s model is better than YOLOv3-tiny, YOLOv5s, YOLOX-s, YOLOX-m, YOLOv6-S, YOLOv6-M, YOLOv7-tiny, YOLOv8s, and YOLOv8m, except YOLOv7, which outperforms the PGDS-YOLOv8s model on mAP@0.5. In addition, the UAV-360 dataset is an equirectangular projected image converted from a fisheye image captured by a fisheye camera mounted on the UAV. On the UAV-360 dataset, the proposed method performs better in the object detection of fisheye images than advanced methods for processing fisheye images, such as LCat, SCat, SLCat, ECDAB, and SphereConv. Among them, the C2fs and SPPFSE modules make the model pay more attention to the information channel features and obtain the best detection accuracy. Meanwhile, to further validate the broad applicability of the proposed method, it is experimented on the MS-COCO 2017 dataset. Compared with the original YOLOv8s model, the PGDS-YOLOv8s model improves detection accuracy.

Limitations. The MPGD and APGD modules effectively extract distorted and incomplete features of fisheye images during the downsampling stage. However, compared with the original YOLOv8s model, the proposed model, with the introduction of the MPGD and APGD modules, takes up slightly more memory during training, which affects the detection speed.

6. Conclusions

In this paper, the proposed PGDS-YOLOv8s model effectively improves the performance of distorted and discontinuous object detection for fisheye images. Firstly, two novel downsampling modules are proposed to demonstrate the superiority of detecting distorted and discontinuous objects. Among them, the MPGD module effectively extracts essential feature information about the distorted and discontinuous objects, and the APGD module can acquire the global field of view and reduce the feature loss of distorted and discontinuous objects. The two downsampling modules expand the sensing field to acquire more feature information and to construct a more lightweight network. Furthermore, the proposed C2fs module acquires richer gradient flow information about features through the interdependence of SE block modeling channels. Meanwhile, the SPPFSE module is an SE module added after the SPPF module. The SPPFSE module can perceive the rich semantic information more efficiently to improve the whole model's ability to detect distorted and discontinuous objects in fisheye images.

On the VOC-360 dataset, the proposed PGDS-YOLOv8s model obtains the best detection compared with YOLOv3-tiny, YOLOv5s, YOLOX-s, YOLOX-m, YOLOv6-S, YOLOv6-M, YOLOv7-tiny, and YOLOv8m. In addition, the improved model on the UAV-360 dataset achieves 89.0% for mAP@ 0.5 and 60.5% for mAP@ 0.5:0.95. Meanwhile, the detection results of our proposed method on the UAV-360 dataset all outperform other object detection methods for fisheye images. Furthermore, the PGDS-YOLOv8s model is used on the MS-COCO 2017 dataset for detection, and the AP is improved by 1.4%, AP_{50} by 1.7%, AP_{75} by 1.2%, AP_S by 2.5%, AP_M by 1.1%, and AP_L by 1.7%. The proposed model also works well for conventional image detection, which proves its broad capability. In the future of

computer vision, the increasing need for wide-angle imaging means that fisheye cameras will be widely studied and used. Therefore, the next part of this paper is dedicated to further investigating how to improve the lightweight characteristics of the model and the performance of the fisheye camera for object detection in real application scenarios.

Author Contributions: Conceptualization, D.Y., J.Z. and T.S.; methodology, D.Y., J.Z. and T.S.; software, J.Z., D.Y. and T.S.; validation, J.Z., D.Y. and T.S.; formal analysis, D.Y., J.Z. and T.S.; investigation, D.Y., J.Z. and T.S.; resources, D.Y. and T.S.; data curation, J.Z., X.Z. and Y.S.; writing—original draft preparation, D.Y., J.Z. and T.S.; writing—review and editing, D.Y., J.Z. and T.S.; visualization, J.Z., X.Z. and Y.S.; supervision, D.Y. and T.S.; project administration, D.Y. and T.S.; funding acquisition, D.Y. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by Natural Science Foundation of Chongqing (CSTB 2022NSCQ-MSX1200), in part by the Science and Technology Research Program of Chongqing Municipal Education Commission (KJZD-M202300502 and KJQN202200537), and in part by Chongqing Normal University Ph.D. Start-up Fund (21XLB035).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study were obtained from [39].

Acknowledgments: The authors thank the editors and anonymous reviewers for their critical comments and suggestions for improving the manuscript.

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

References

1. Song, J.; Yu, Z.; Qi, G.; Su, Q.; Xie, J.; Liu, W. UAV Image Small Object Detection Based on RSAD Algorithm. *Appl. Sci.* **2023**, *13*, 11524. [CrossRef]
2. Mou, C.; Liu, T.; Zhu, C.; Cui, X. WAID: A Large-Scale Dataset for Wildlife Detection with Drones. *Appl. Sci.* **2023**, *13*, 10397. [CrossRef]
3. Barmpoutis, P.; Stathaki, T.; Dimitropoulos, K.; Grammalidis, N. Early fire detection based on aerial 360-degree sensors, deep convolution neural networks and exploitation of fire dynamic textures. *Remote Sens.* **2020**, *12*, 3177. [CrossRef]
4. Luo, C.; Yu, L.; Yan, J.; Li, Z.; Ren, P.; Bai, X.; Yang, E.; Liu, Y. Autonomous detection of damage to multiple steel surfaces from 360 panoramas using deep neural networks. *Comput. Aided Civ. Infrastruct. Eng.* **2021**, *36*, 1585–1599. [CrossRef]
5. Gao, W.; Wang, K.; Ding, W.; Gao, F.; Qin, T.; Shen, S. Autonomous aerial robot using dual-fisheye cameras. *J. Field Robot.* **2020**, *37*, 497–514. [CrossRef]
6. Yang, T.; Ren, Q.; Zhang, F.; Xie, B.; Ren, H.; Li, J.; Zhang, Y. Hybrid Camera Array-Based UAV Auto-Landing on Moving UGV in GPS-Denied Environment. *Remote Sens.* **2018**, *10*, 1829. [CrossRef]
7. Kumar, V.R.; Yogamani, S.; Rashed, H.; Sitsu, G.; Witt, C.; Leang, I.; Milz, S.; Mäder, P. Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving. *IEEE Robot. Autom. Lett.* **2021**, *6*, 2830–2837. [CrossRef]
8. Cui, Z.; Heng, L.; Yeo, Y.C.; Geiger, A.; Pollefeys, M.; Sattler, T. Real-time dense mapping for self-driving vehicles using fisheye cameras. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 6087–6093. [CrossRef]
9. Billings, G.; Johnson-Roberson, M. SilhoNet-fisheye: Adaptation of a ROI based object pose estimation network to monocular fisheye images. *IEEE Robot. Autom. Lett.* **2020**, *5*, 4241–4248. [CrossRef]
10. Roxas, M.; Oishi, T. Variational fisheye stereo. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1303–1310. [CrossRef]
11. Benseddik, H.E.; Morbidi, F.; Caron, G. PanoraMIS: An ultra-wide field of view image dataset for vision-based robot-motion estimation. *Int. J. Robot. Res.* **2020**, *39*, 1037–1051. [CrossRef]
12. Itakura, K.; Hosoi, F. Automatic Tree Detection from Three-Dimensional Images Reconstructed from 360° Spherical Camera Using YOLO v2. *Remote Sens.* **2020**, *12*, 988. [CrossRef]
13. Yang, L.; Hu, G.; Song, Y.; Li, G.; Xie, L. Intelligent video analysis: A Pedestrian trajectory extraction method for the whole indoor space without blind areas. *Comput. Vis. Image Underst.* **2020**, *196*, 102968. [CrossRef]
14. Bertel, T.; Yuan, M.; Lindroos, R.; Richardt, C. Omniphotos: Casual 360 vr photography. *ACM Trans. Graph. TOG* **2020**, *39*, 1–12. [CrossRef]
15. Zhou, Y.; Tian, L.; Zhu, C.; Jin, X.; Sun, Y. Video coding optimization for virtual reality 360-degree source. *IEEE J. Sel. Top. Signal Process.* **2019**, *14*, 118–129. [CrossRef]

16. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
17. Mao, J.; Niu, M.; Jiang, C.; Liang, H.; Chen, J.; Liang, X.; Li, Y.; Ye, C.; Zhang, W.; Li, Z.; et al. One million scenes for autonomous driving: Once dataset. *arXiv* **2021**, arXiv:2106.11037. [CrossRef]
18. Naude, J.; Joubert, D. The Aerial Elephant Dataset: A New Public Benchmark for Aerial Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 48–55.
19. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 year, 1000 km: The oxford robotcar dataset. *Int. J. Robot. Res.* **2017**, *36*, 3–15. [CrossRef]
20. Yogamani, S.; Hughes, C.; Horgan, J.; Sistu, G.; Varley, P.; O'Dea, D.; Uricár, M.; Milz, S.; Simon, M.; Amende, K.; et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 16–17 June 2019; pp. 9308–9318.
21. Chiang, S.H.; Wang, T.; Chen, Y.F. Efficient pedestrian detection in top-view fisheye images using compositions of perspective view patches. *Image Vis. Comput.* **2021**, *105*, 104069. [CrossRef]
22. Chen, P.Y.; Hsieh, J.W.; Gochoo, M.; Wang, C.Y.; Liao, H.Y.M. Smaller object detection for real-time embedded traffic flow estimation using fish-eye cameras. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2956–2960. [CrossRef]
23. Arsenali, B.; Viswanath, P.; Novosel, J. RotInvMTL: Rotation invariant MultiNet on fisheye images for autonomous driving applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019. [CrossRef]
24. Wei, X.; Wei, Y.; Lu, X. RMDC: Rotation-mask deformable convolution for object detection in top-view fisheye cameras. *Neurocomputing* **2022**, *504*, 99–108. [CrossRef]
25. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
26. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
27. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [CrossRef]
28. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
29. 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. [CrossRef]
30. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271. [CrossRef]
31. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
32. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
33. Glenn, J. YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 5 July 2023).
34. 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. [CrossRef]
35. 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. *arXiv* **2022**, arXiv:2207.02696. [CrossRef]
36. Glenn, J. YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 10 March 2023).
37. Ju, R.Y.; Cai, W. Fracture Detection in Pediatric Wrist Trauma X-ray Images Using YOLOv8 Algorithm. *arXiv* **2023**, arXiv:2304.05071. [CrossRef]
38. Zhai, X.; Huang, Z.; Li, T.; Liu, H.; Wang, S. YOLO-Drone: An Optimized YOLOv8 Network for Tiny UAV Object Detection. *Electronics* **2023**, *12*, 3664. [CrossRef]
39. Fu, J.; Bajić, I.V.; Vaughan, R.G. Datasets for face and object detection in fisheye images. *Data Brief* **2019**, *27*, 104752. [CrossRef] [PubMed]
40. Williams, T.; Li, R. Wavelet Pooling for Convolutional Neural Networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://openreview.net/forum?id=rkhlb8lCZ (accessed on 17 November 2023).
41. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features From Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [CrossRef]
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]

43. Liu, C.; Yang, D.; Tang, L.; Zhou, X.; Deng, Y. A lightweight object detector based on spatial-coordinate self-attention for UAV aerial images. *Remote Sens.* **2023**, *15*, 83. [CrossRef]
44. Kim, S.; Park, S.Y. Expandable Spherical Projection and Feature Concatenation Methods for Real-Time Road Object Detection Using Fisheye Image. *Appl. Sci.* **2022**, *12*, 2403. [CrossRef]
45. Zhang, X.; Yang, D.; Song, T.; Ye, Y.; Zhou, J.; Song, Y. Classification and Object Detection of 360° Omnidirectional Images Based on Continuity-Distortion Processing and Attention Mechanism. *Appl. Sci.* **2022**, *12*, 12398. [CrossRef]
46. Coors, B.; Condurache, A.P.; Geiger, A. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 518–533. [CrossRef]