electronics

Article

ADYOLOv5-Face: An Enhanced YOLO-Based Face Detector for Small Target Faces

Linrunjia Liu [1], Gaoshuai Wang [2] and Qiguang Miao [1,*]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China; liulinrunjia@xidian.edu.cn
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China; wanggaoshuai@sz.tsinghua.edu.cn
* Correspondence: qgmiao@xidian.edu.cn

Abstract: Benefiting from advancements in generic object detectors, significant progress has been achieved in the field of face detection. Among these algorithms, the You Only Look Once (YOLO) series plays an important role due to its low training computation cost. However, we have observed that face detectors based on lightweight YOLO models struggle with accurately detecting small faces. This is because they preserve more semantic information for large faces while compromising the detailed information for small faces. To address this issue, this study makes two contributions to enhance detection performance, particularly for small faces: (1) modifying the neck part of the architecture by integrating a Gather-and-Distribute mechanism instead of the traditional Feature Pyramid Network to tackle the information fusion challenges inherent in YOLO-based models; and (2) incorporating an additional detection head specifically designed for detecting small faces. To evaluate the performance of the proposed face detector, we introduce a new dataset named XD-Face for the face detection task. In the experimental section, the proposed model is trained using the Wider Face dataset and evaluated on both Wider Face and XD-face datasets. Experimental results demonstrate that the proposed face detector outperforms other excellent face detectors across all datasets involving small faces and achieved improvements of 1.1%, 1.09%, and 1.35% in the AP50 metric on the WiderFace validation dataset compared to the baseline YOLOv5s-based face detector.

Keywords: object detection; small face detection; YOLOv5; face dataset; classroom scenes; network structure design

check for updates

Citation: Liu, L.; Wang, G.; Miao, Q. ADYOLOv5-Face: An Enhanced YOLO-Based Face Detector for Small Target Faces. *Electronics* **2024**, *13*, 4184. https://doi.org/10.3390/electronics13214184

Academic Editor: Manohar Das

Received: 13 September 2024
Revised: 12 October 2024
Accepted: 21 October 2024
Published: 25 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Face detection aims to identify and locate faces within images or video streams, serving as a fundamental step in various face-related applications such as face verification, recognition, and expression analysis. With the advent of Convolutional Neural Network (CNN)-based object detectors, face detection has witnessed significant advancements in recent years. Evolving from two-stage approaches like Faster RCNN [1] and Mask RCNN [2] to one-stage methods such as SSD [3], RetinaNet [4] and YOLO [5], the object detection framework has facilitated the development of various face detectors based on these methods. Notable examples include RetinaFace [6], YOLOv5Face [7], RefineFace [8], and TinaFace [9]. Among these, YOLO series detectors stand out due to their low computational cost and high accuracy, making them widely adopted in industry applications.

However, despite the effectiveness of those deep learning-based models, existing face detectors exhibit a significant drawback: while they excel in recognizing large faces, their accuracy diminishes when it comes to detecting tiny faces. This disparity arises because large-scale targets necessitate a larger receptive field, causing the features of small targets to gradually fade on the deep feature map after multiple downsampling steps.

To address this issue, FaceBoxes [10] designed multi-scale anchors to enrich the receptive field and discretized anchors across different layers to handle faces of various sizes.

YOLOFaceV2 introduced a novel Receptive Field Enhancement module (RFE), providing a richer receptive field to capture both large and small-scale faces simultaneously. Reference [11] explored using attention mechanisms for facial feature extraction, preventing the misidentification of small-scale face features due to occlusion. However, these methods failed to fully utilize information between channels. They utilize a Feature Pyramid Network (FPN) in the neck architecture, which integrates information layer by layer, implying that features from different layers cannot be fused directly without passing through intermediate layers. Inspired by Gold-YOLO [12], we integrate a novel Gather-and-Distribute (GD) mechanism in the neck architecture, replacing FPN. This modification allows for the seamless transmission of information across layers, thereby enhancing feature fusion capabilities.

Other research methods improved small-face detection by deploying a large number of small anchors across the image. While these methods effectively increased the recall rate, they also exacerbated the issue of extreme class imbalance, leading to numerous false positives in detection results [13]. To address this problem, Lin et al. introduced Focal Loss, which dynamically adjusts the weight of hard samples [4]. Similarly, the Gradient Harmonizing Mechanism (GHM) [14] suppresses the gradients of easy samples to prioritize hard ones, while PISA [15] assigns weights to positive and negative samples based on different criteria. YOLOFaceV2 [16] proposed a Slide Weight Function (SWF) to adjust the weights of imbalanced samples, enabling adaptive learning of the threshold parameters for positive and negative samples. Building on previous work, we incorporate four detection heads instead of three, for detecting tiny, small, medium, and large faces. While the additional tiny head increases small anchors for detecting small objects, Normalized Wasserstein Distance (NWD) loss and Intersection over Union (IOU) loss are combined in this work to balance the detection of both large and small samples.

To evaluate the performance of our proposed method, we create a dataset specifically focusing on scenes with small faces, named the XD-face dataset. We choose 2802 images and label 102,250 faces with small size and occlusion in classroom scenes. We compare the recognition results of our proposed model on both the WiderFace and XD-face datasets with those of other state-of-the-art approaches. With our self-trained model, our method achieves mAP scores of 94.80%, 93.77%, and 84.37% on the easy, medium, and hard subsets of the Wider Face dataset, respectively, and 59.7% on the XD-face dataset.

The key contributions of this work are summarized as follows:

(1) Improvement of the neck structure of YOLO series models and utilization of the GD mechanism instead of the FPN to enhance information fusion capabilities;

(2) Addition of an extra detection head to improve the detection of tiny faces;

(3) Constructing a novel dataset for face detection, focusing specifically on classroom scenes with small faces.

2. Related Work

2.1. Two-Stage Face Detection Methods

In recent years, face detection methods based on convolutional neural networks can generally be divided into two-stage methods and single-stage methods. The core idea of two-stage face detection methods is to generate a limited number of candidate regions (Proposals) in the first stage, and then refine the candidate set in the second stage to generate the final results.

Faster R-CNN [1] is a typical two-stage object detection method. In the first stage, it utilizes a Region Proposal Network (RPN) to generate a series of candidate regions, followed by classification and regression predictions using candidate boxes in the second stage. Jiang et al. [17] first applied the Faster R-CNN model to face detection and achieved good detection results on two widely used face detection datasets. Sun et al. [18] improved upon Faster R-CNN by employing feature concatenation, hard sample mining, multi-scale training, model pre-training, and fine-tuning of key parameters to enhance face detection performance. Zhu et al. [19] subsequently proposed CMS-RCNN, which detects

unconstrained faces by utilizing contextual information. Khan et al. [20] used multi-task cascaded convolutional neural networks and modified the layer density with increasing neuron count. It enriches feature information while increasing the amount of calculation. Nonetheless, early methods of face detection based on deep learning exhibited significant drawbacks, including intricate training processes, a susceptibility to local optima, protracted detection times, and suboptimal accuracy.

2.2. One-Stage Face Detection Methods

Due to the substantial computational time required by two-stage face detection methods, single-stage face detection methods were subsequently proposed. Single-stage face detection methods based on SSD [3] and YOLO [5] do not require generating candidate regions.

Compared to two-stage methods, single-stage methods offer faster detection speed and represent a type of end-to-end face detection method. The network structure of single-stage face detection methods can generally be divided into three parts: backbone network, neck network, and head network.

Backbone: The backbone network usually employs models such as VGG [21], ResNet [22], DenseNet [23,24], SENet [25], MobileNet [26], ShuffleNet [27], etc. BFBox [28] searches for backbone networks suitable for face space. SCRFD [29] introduces a computation redistribution method to redistribute computational resources between the backbone network, neck network, and head network of the model.

Neck: After the introduction of the FPN structure [30], it often serves as the basic neck network for face detection. FANet [31] creates a new hierarchical efficient FPN with rich semantics at all scales. BFBox [28] proposes an FPN attention module to jointly search the architecture of the backbone network and FPN.

To further enhance face features, some face detection methods choose to use different convolution kernels in each branch to enlarge the receptive field. SSH [32] adds larger convolution kernels to each detection module to incorporate contextual information. Wang et al. [33] conducted research on low-light face detection in unannotated scenarios by combining context learning with contrastive learning. DSFD [34] introduces a feature enhancement module to strengthen the original features, making them more recognizable and robust. RefineFace [8] constructs an RFE module to provide diversified perspectives for detecting faces with extreme poses. SmallHardFace [35] expands the receptive field using dilated convolution. YOLO5Face [7] introduces face keypoint loss into the network and improves the SPP module in YOLOv5 by using smaller convolution kernels instead of larger ones to make it more suitable for face detection.

Although these methods enhance the receptive field and enrich semantic information, providing a simple and effective means to improve multi-scale object detection performance, they may result in the loss of spatial information. Therefore, this paper proposes the use of a GD mechanism instead of FPN in the neck architecture to combine deep semantic information with shallow features.

Head: The head part of the network structure connects the loss function for optimizing model parameters through backpropagation. RetinaFace [6] introduced facial keypoint information into the face detection system, thereby augmenting the facial keypoint regression loss function. Fang et al. [36] introduced the hierarchical loss Triple Loss to optimize the face detector based on the feature pyramid network. Inspired by IoU-awareness [37], TinaFace [9] incorporated a regression branch into the head network and utilized cross-entropy loss to predict the IoU value between annotated bounding boxes and detection results. Although the introduction of different loss functions alleviates the issue of sample imbalance, it also introduces additional hyperparameters, making the model training more complex.

Considering some of the current advanced face detection algorithms are designed to be overly complex and have high computational costs, this paper proposes improvements based on the commonly used object detection model YOLOv5. To solve the information fusion problem of the traditional FPN structure, this paper integrates a GD mechanism.

For improving the detection accuracy of small faces, a tiny-face detection head is added in this paper.

3. Proposed Method

3.1. Overview of ADYOLOv5-Face

With the maturity of facial detection neural network technology, most one-stage facial detection neural models possess a three-stage structure consisting of backbone, neck, and detection networks. The backbone network is responsible for extracting high-dimensional feature information: locational information on high-resolution feature maps and semantic information on low-resolution feature maps. Then, the neck network fuses different features extracted by the backbone network. Finally, the head network usually contains three detection heads for detecting small, medium and large objects.

This paper aims to focus on the information fusion capability of the neck network and introduce the GD mechanism to resolve the issue of information loss caused by cross-layer feature integration in current models. In the head part, one more head for tiny object detection is added. As for the backbone part, the CSPDarknet53 was kept unchanged, identical to its structure in YOLOv5. The architecture of our ADYOLOv5-Face detector is depicted in Figure 1.

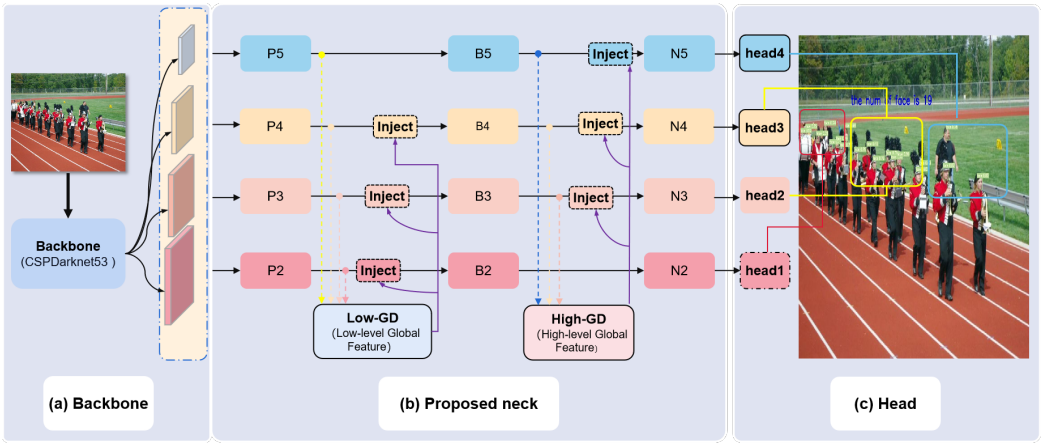


Figure 1. The architecture of the ADYOLOv5-Face: (**a**) CSPDarknet53 backbone as the original YOLOv5 version. (**b**) The neck uses a structure like the Gather-and-Distribute mechanism. (**c**) Four prediction heads use the feature maps from the neck. Head1 contains more detailed information that is added to detect tiny faces.

3.2. Architecture of the Neck Part

The backbone part is responsible for extracting high-dimensional feature information, while the neck part fuses different features extracted by the backbone network—namely, the positional information from high-resolution feature maps and the semantic information from low-resolution feature maps. As shown in Figure 2a, the proposed neck structure in this project is as follows:

Step 1: Concatenate the four layers of features, $P2$, $P3$, $P4$, and $P5$, extracted by the backbone network to obtain the low-level global feature F_{global_low}.

Step 2: Inject the low-level global feature F_{global_low} back into the local features $P2$, $P3$, and $P4$, achieving an initial fusion of global and local features, resulting in features $B2$, $B3$, and $B4$. $B2$ will be directly used as the final output, preserving the high-resolution positional information of the low-level features without loss.

Step 3: The low-level fused features $B2$, $B3$, and $B4$ are further concatenated to obtain a high-level global feature F_{global_high} with richer semantic information. To fully integrate the global features, a transformer module is proposed, combining attention mechanisms and convolution operations to better handle information from different positions in the image.

Step 4: Finally, $F2$ is injected back into the $B3$, $B4$, and $B5$ features, resulting in high-level fused features $N3$, $N4$, and $N5$, which carry rich semantic information and are fed into the prediction network as output layer features.

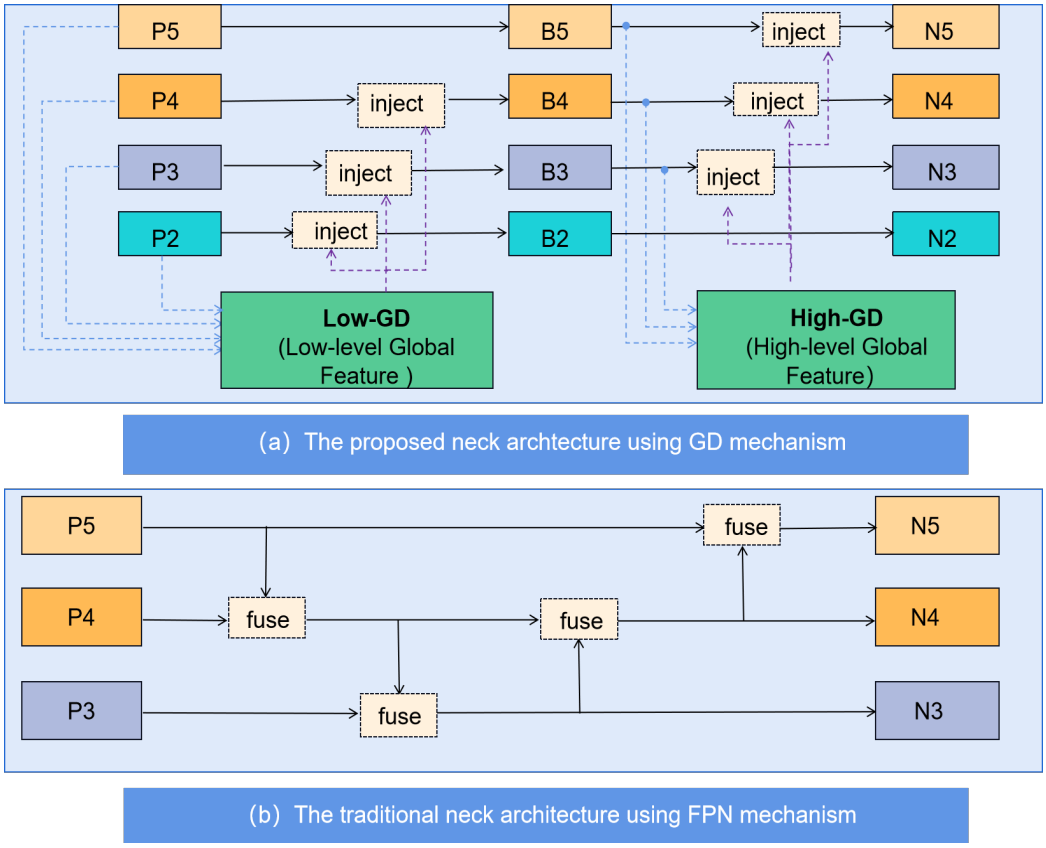


Figure 2. The comparison of the proposed neck part with the traditional neck part.

Compared to the traditional mid-level network structure shown in Figure 2b, the complex mid-level network proposed in this paper directly fuses global features with features from each layer during feature fusion, avoiding information loss that may occur during cross-layer transmission and thus enhancing small object detection performance. At the same time, the design of the global feature fusion module also considers the integration of new technologies with traditional methods, improving the model's feature fusion capability without significantly increasing latency, thereby enhancing the object detection performance. Lastly, since the focus of this project is on detecting small faces, the model strengthens the use of low-level information ($P2$) by adding a low-level output feature ($N2$) to preserve more detailed features.

3.3. Details of the Neck Part GD Mechanism

The GD mechanism consists of three parts: feature alignment module (FAM), information fusion module (IFM) and information injection module (IIM).

To better represent the features of varying sizes, the GD mechanism can be divided into low-stage and high-stage. The feature alignment module and information fusion module are different in the two GD mechanism branches, while the information injection module is the same.

- Low-stage Gather-and-Distribute mechanism

For the low-stage Gather-and-Distribute mechanism (Low-GD), the structure is shown in Figure 3. The details are provided as follows:

(1) Align the four features $P2$, $P3$, $P4$, and $P5$ extracted by the backbone network in a unified size using the average pooling operation (AvgPool) and Bilinear operation.

$$F_{align} = Low_FAM([P2, P3, P4, P5]) \tag{1}$$

(2) Concatenate the four features, and employ RepBlock to obtain the low-level global feature.

$$F_{fuse} = RepBlock(F_{align}) \tag{2}$$

(3) Split the low-level global feature into F_{inj_B2}, F_{inj_B3} and F_{inj_B4} with different channel dimensions.

$$F_{global_low} = F_{inj_B2}, F_{inj_B3}, F_{inj_B4} = Split(F_{fuse}) \tag{3}$$

(4) Next is the information injection module. As shown in Figure 1, the three split features (F_{inj_Bi}) are injected into the local features $P2$, $P3$, and $P4$ to achieve the initial fusion of global and local features, resulting in features $B2$, $B3$, and $B4$. $B2$ will be directly output as the final feature, preserving the locational information of the low-level high-resolution features without loss. The details of the information injection module are shown in Figure 4, and the formula is as follows:

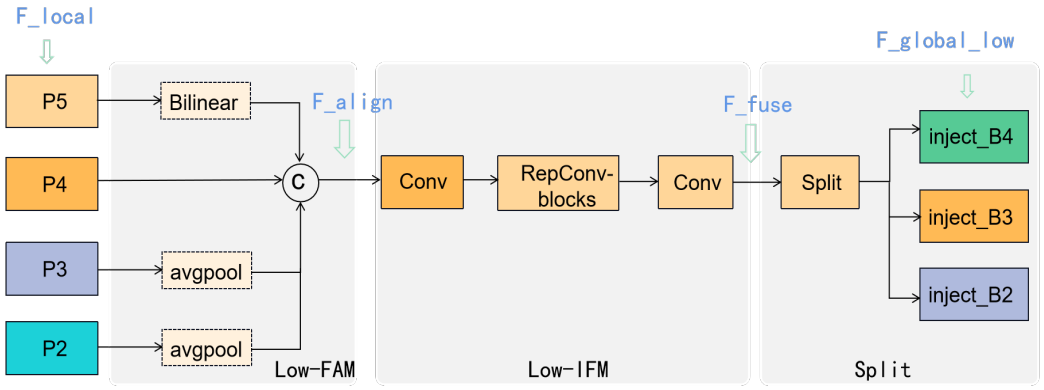


Figure 3. The structure of the Low-GD in the neck part. The notations in blue are aligned with the corresponding features in Equations (1)–(3).

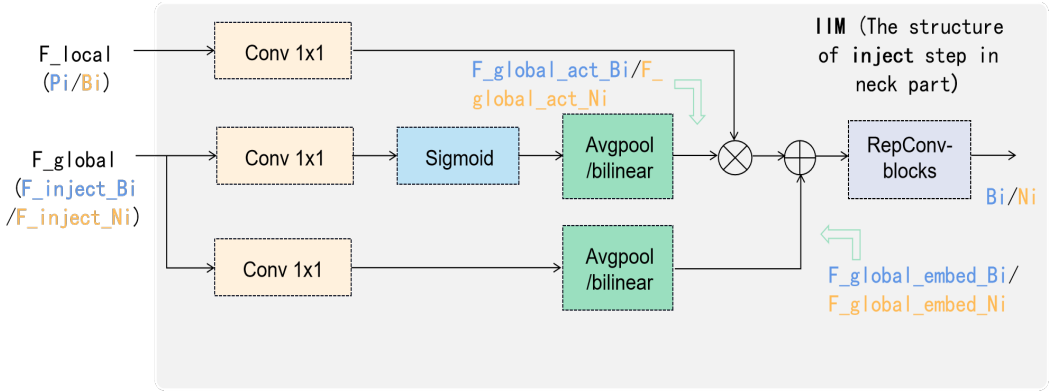


Figure 4. The structure of inject step in the neck part, which is named the information injection module (IIM). The notations in blue are aligned with the corresponding features in Equations (4)–(7), while the notations in yellow are aligned with the corresponding features in Equations (11)–(14).

$$F_{global_act_Bi} = resize(sigmoid(Conv(F_{inj_Bi}))) \tag{4}$$

$$F_{global_embed_Bi} = resize(Conv(F_{inj_Bi})) \tag{5}$$

$$F_{att_fuse_Bi} = Conv(Pi) * F_{global_act_Bi} + F_{global_embed_Bi} \tag{6}$$

$$Bi = RepBlock(F_{att_fuse_Bi}) \tag{7}$$

- High-stage gather-and-distribute mechanism

For the high-stage Gather-and-Gistribute mechanism (High-GD), the structure is shown in Figure 5. The low-order fused features $B3$, $B4$, and $B5$ will be further concatenated to obtain the high-order global feature with richer semantic information. In order to fully integrate global features, a transformer module instead of the RepConv-Block is proposed to better handle information from different positions in the image by combining attention mechanisms with convolution operations.

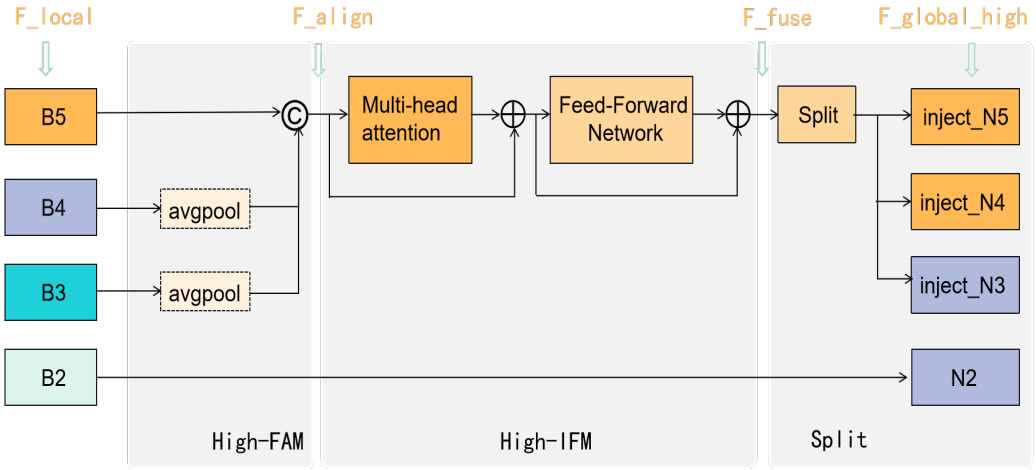


Figure 5. The structure of the High-GD in the neck part. The notations in blue are aligned with the corresponding features in Equations (8)–(10).

The details are provided as follows:

(1) Align the three features $B3$, $B4$, and $B5$ extracted by the backbone network in a unified size by AvgPool.

$$F_{align} = High_FAM([B3, B4, B5]) \tag{8}$$

(2) Concatenate the three features and employ the transformer fusion module to obtain the high-level global feature.

$$F_{fuse} = Transformer(F_{align}) \tag{9}$$

(3) Split the high-level global feature into F_{inj_N3}, F_{inj_N4} and F_{inj_N5} with different channel dimensions.

$$F_{global_high} = F_{inj_N3}, F_{inj_N4}, F_{inj_N5} = Split(Conv1 \times 1(F_{fuse})) \tag{10}$$

(4) The information injection module here is the same as in the Low-GD branch, as shown in Figure 4. The three split features (F_{inj_Ni}) are injected into the features $B3$, $B4$, and $B5$ to obtain high-order fused information $N3$, $N4$, and $N5$ with rich semantic information, which are then fed into the detection network as output layer features. The formula is as follows:

$$F_{global_act_Ni} = resize(sigmoid(Conv(F_{inj_Ni}))) \tag{11}$$

$$F_{global_embed_Ni} = resize(Conv(F_{inj_Ni})) \tag{12}$$

$$F_{att_fuse_Ni} = Conv(Bi) * F_{global_act_Ni} + F_{global_embed_Ni} \tag{13}$$

$$Ni = RepBlock(F_{att_fuse_Ni}) \tag{14}$$

3.4. Prediction Head for Tiny Faces

After investigating the WiderFace dataset, we find that it contains many extremely small faces. Therefore, one more head is added to the proposed ADYOLOv5-Face for detecting tiny faces. As shown in Figure 1, head1 is generated from feature map *B*2, which contains more details of the tiny faces compared to the other three prediction heads. The performance of tiny face detection has large improvements after adding the tiny prediction head, even though the computation cost has also increased.

3.5. Loss

Since metrics based on IoU, such as IoU itself and its extensions, are highly sensitive to positional deviations of small objects, ref. [38] proposed a new evaluation method for small object detection based on the Wasserstein distance. Equation (15) calculates the Wasserstein distance between the predicted object and the ground truth object.

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[cx_b, cy_b, \frac{w_b}{2} \cdot \frac{h_b}{2} \right]^T \right) \right\|_2^2 \tag{15}$$

where c is a constant specific to the dataset, $\mathcal{N}_a$ and $\mathcal{N}_b$ are Gaussian distributions characterized by bounding boxes $A = (cx_a, cy_a, w_a, h_a)$ and $B = (cx_b, cy_b, w_b, h_b)$, respectively.

However, $W_2^2(\mathcal{N}_a, \mathcal{N}_b)$ is a distance metric and cannot be directly used as a similarity measure (i.e., a value between 0 and 1 like IoU). Therefore, we normalize its exponential form and obtain a new metric called the Normalized Wasserstein Distance (NWD), as shown in Equation (16).

$$NWD(\mathcal{N}_a, \mathcal{N}_b) = exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{C} \right) \tag{16}$$

The loss function based on the NWD metric is expressed as follows:

$$\mathcal{L}(NWD) = 1 - NWD(\mathcal{N}_a, \mathcal{N}_b) \tag{17}$$

Therefore, this paper uses the NWD loss to compensate for the shortcomings of IoU loss in small object detection while retaining the IoU loss for detecting large objects.

4. Experiment Setup

4.1. Datasets

Proposed Dataset (XD-face): The XD-face dataset comprises face detection data extracted from nine classroom videos captured within our school environment, encompassing a wide array of face images varying in size, pose, and occlusion levels. As illustrated in Table 1, the dataset consists of 2802 images and a total of 102,250 labeled faces, each annotated with high accuracy (with data organized as 5 values per row: category, x_center, y_center, width, height), reflecting variations in illumination and face density. Due to the fixed camera positioning within the classroom setting, even in scenes with few individuals, the faces captured tend to be very small. The dataset meets the requirements in this paper for detecting small faces and is thus utilized to evaluate the proposed model, benchmarked against other YOLO-based methods. Figure 6 illustrates the visualization of the conditions and detection results in XD-face. Even though the faces in this dataset are very small and are occluded in most locations, they are detected by our detector.

The XD-face dataset, as a classroom scene dataset, holds potential for applications in face counting, face recognition, and domain adaptation.

Table 1. The contents of the proposed XD-face dataset encompass a rich array of facial images in different scenes.

Name	Images	Faces	Dense/Bright	Dense/Dark	Sparse/Bright	Sparse/Dark
A-208	142	6208	142	0	0	0
A-211	150	2001	100	50	0	0
B-301	320	25,108	160	160	0	0
B-403	300	11,775	150	150	0	0
B-418	315	17,351	165	150	0	0
B-507	335	21,478	185	150	0	0
B-509	300	2216	0	0	0	300
C-107	400	5373	0	0	400	0
C-225	540	10,740	300	0	240	0
Total	2802	102,250	1202	660	640	300



Figure 6. Visual comparisons of detecting conditions and results in our dataset: XD-face. XD-face contains 2802 images and labels 102,250 faces with small size and occlusion in several different classroom scenes.

Firstly, in terms of face counting, the diverse classroom scenes in the dataset can assist researchers in developing more accurate algorithms for identifying and counting the number of students in a class, which is crucial for educational management and classroom optimization.

In face recognition, the XD-face dataset can enhance the efficiency of classroom attendance and identity verification systems, such as automatically recognizing students and recording attendance. Additionally, teachers can use this technology to track student engagement and performance, providing data to support personalized teaching.

Regarding domain adaptation, the features of this dataset allow algorithms to adjust across different classroom environments, improving recognition under varying lighting

conditions and angles. This is vital for ensuring system stability and reliability in remote or hybrid learning environments.

Overall, the XD-face dataset provides a rich foundation for the application of facial recognition technologies in education, contributing to improved teaching efficiency and student management.

Wider Face: As a widely recognized benchmark dataset utilized across different face detection methods, Wider Face comprises 61 scene classes, totaling 32,203 images and 393,703 labeled faces. Distinguished from the proposed XD-face dataset, Wider Face encompasses a diverse range of facial sizes, including small, medium, and large faces. The dataset categorizes faces into three levels based on their sizes: Easy, Medium, and Hard, with the Hard level representing the most challenging subset, ideal for evaluating the accuracy of proposed models. In this study, the training set, comprising 40% of the dataset, is utilized for training the proposed model, while the validation set, comprising 10% of the dataset, is employed for evaluation purposes. The evaluation metric employed for Wider Face is the mean average precision at an IOU threshold of 0.5 (mAP@.5).

4.2. Experimental Evaluation Metrics

This paper selects Average Precision (AP) [39] as the evaluation metric. The area under the PR curve represents the AP.

PR Curve: It reflects the relationship between precision and recall. The horizontal axis represents recall, and the vertical axis represents precision.

Confusion Matrix: A confusion matrix is a crucial tool for evaluating the performance of classification models, especially for binary and multi-class classification tasks. For binary classification problems, the confusion matrix typically takes a 2×2 form. The confusion matrix is defined as shown in Table 2:

Table 2. Confusion matrix for face detection model.

Confusion Matrix		**Actual**	
		Face	**Non-Face**
Predicted	Face	*TP*	*FP*
	Non-Face	*FN*	*TN*

- True Positive (TP): the number of instances correctly predicted as positive (face).
- False Negative (FN): the number of instances incorrectly predicted as negative (non-face) when they are actually positive (face).
- False Positive (FP): the number of instances incorrectly predicted as positive (face) when they are actually negative (non-face).
- True Negative (TN): the number of instances correctly predicted as negative (non-face).

Precision measures the proportion of true positive instances among the instances predicted as positive by the model, indicating the accuracy of the model's positive predictions. The formula for precision is given in Equation (18):

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

Recall measures the proportion of true positive instances among all actual positive instances, indicating the model's coverage ability. The formula for recall is given in Equation (19):

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

In face detection, a key metric for evaluating detection results is IoU, which measures the overlap between the predicted bounding box and the ground truth bounding box. It

is an important standard for assessing model accuracy. The IoU is calculated using the formula shown in Equation (20):

$$IoU = \frac{A \cap B}{A \cup B} \tag{20}$$

Here, B is the predicted bounding box and G is the ground truth bounding box. If the IoU of the predicted box and the ground truth box is greater than a preset threshold (usually 0.5), the predicted box is considered a TP; otherwise, it is considered an FP. Precision and recall are then calculated using Equations (18) and (19).

AP is a summary measure of the PR curve, assessing the model's overall performance across different thresholds. AP is the numerical integration of the area under the PR curve, obtained by calculating the average precision across all recall levels. The higher the AP, the better the model performs across different recall levels.

AP50: IOU for the NMS is set to 0.5.

AP@50:5:95: The IOU values range from 0.5 to 0.95, with a step size of 0.5.

Since AP50 is a commonly used metric in object detection, we primarily assess the performance of our detector and other state-of-the-art face detectors using AP50. For the experiments on XD-face, however, we evaluate the performance of these detectors across four metrics—precision, recall, AP50, and AP50:5:95—in order to highlight the detection difficulty on the proposed dataset.

4.3. Ablation Experiment

We utilize the YOLOv5s model as our baseline and integrate the occlusion-aware repulsion loss, drawing inspiration from the work of [16]. The training process runs 250 epochs on the Wider Face training set with a batch size of 32. For optimization, we employ SGD with momentum. The initial learning rate is set to 1×10^{-2}, gradually decaying to 1×10^{-5}, with a weight decay of 5×10^{-3}. Following this configuration, the proposed ADYOLOv5-Face model implements modifications to the neck and head structures.

The detection accuracy (using AP50 as the metric) of the baseline model and the proposed model in the validation set of Wider Face is shown in Table 3. For the baseline model, the detection accuracy in easy, medium, are hard levels is 93.70%, 92.68% and 83.02%, respectively. When the GD mechanism is applied into the neck structure, the detection accuracy of easy and medium levels is increased by 1.46% and 0.74%, respectively, while the detection performance in the hard level is damaged by 1.66%. To improve the detection accuracy of small faces, a tiny face detection layer is added to enhance the representation ability of detail features. Therefore, the advanced model achieves 94.80%, 93.77% and 84.37% detection accuracy in easy, medium, are hard levels, respectively. Compared to the baseline model, the ADYOLOv5-Face model has 1.1%, 1.09% and 1.35% increments in easy, medium, and hard levels, respectively. In this way, the proposed model improves the detection ability for small target faces without sacrificing the recognition accuracy of large faces.

Table 3. Ablation study results of the AP50 metric on the Wider Face validation set.

Modification	Easy (%)	Medium (%)	Hard (%)	Params (M)	Flops (G)
baseline	93.70	92.68	83.02	7.063	16.4
+ neck	95.16	93.42	81.36	10.008	21.2
+neck +head	94.80	93.77	84.37	10.123	22.8

4.4. Contrast Experiment

4.4.1. Experiments on Wider Face

In this section, we compare our model with various state-of-the-art face detectors, as presented in Table 4, which includes detectors based on ResNet, YOLOv5, YOLOv7 [40], and YOLOv8 networks. The data presented in the table are obtained from the paper [7].

Table 4. Comparison of our ADYOLOv5-Face with the state-of-the-art face detectors on the Wider Face validation dataset (using AP50 as the evaluation metric).

Detector	Backbone	Easy (%)	Medium (%)	Hard (%)
DSFD (2019) [34]	ResNet152	94.29	91.47	71.39
RetinaFace (2020) [6]	ResNet50	94.92	91.90	64.17
HAMBox (2020) [41]	ResNet50	95.27	93.76	76.75
TinaFace (2020) [9]	ResNet50	95.61	94.25	81.43
SCRFD-2.5GF (2021) [29]	Basic ResNet	93.78	92.16	77.87
TinyYolov3 (2022) [42]	YOLOv3-tiny	95.26	89.2	77.9
YOLOv7-tiny-Face (2022) [7]	YOLOv7-tiny	94.7	92.6	82.1
YFaces-Tiny (2024) [43]	YOLOv7-tiny	94.07	92.36	83.06
YOLOv8n-face (2023) [7]	YOLOv8n	94.5	92.2	79
YOLOv5s-Face (2021) [7]	YOLOv5s	94.33	92.61	83.15
ADYOLOv5-Face (ours)	YOLOv5s	94.80	93.77	84.37

Upon evaluating the detection performances (using AP50 as the metric) for large-scale faces, it is observed that TinaFace achieves the highest detection accuracy in the easy (95.61%) and medium (94.25%) subsets on the Wider Face validation dataset, slightly outperforming our proposed model (94.80% and 93.77%, respectively). However, when evaluating the detection performance for small faces, our proposed model outperforms other face detectors. Specifically, our proposed model achieves an 84.3% detection accuracy in the hard-level subset, demonstrating a 20.2% improvement over RetinaFace and 1.22% over the second-best model, YOLOv5s-Face.

In comparison with YOLOv5s-Face, YOLOv7-tiny-Face, and YOLOv8n-face, which belong to the same YOLO algorithm and possess similar model size and computational requirements, our proposed model attains the highest accuracy across all three subsets. Notably, it demonstrates improvements of 1.22%, 2.27%, and 5.37%, respectively, in the hard-level subset. Figure 7 provides a clearer illustration of the superiority of the proposed method's performance on the hard set.

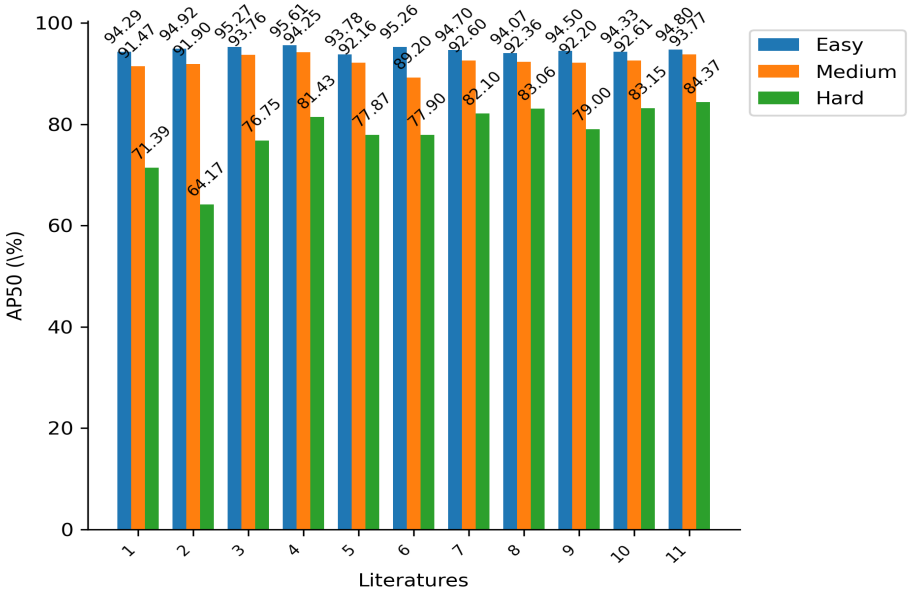


Figure 7. Detection results on the Wider Face validation dataset. Literature 1: DSFD (2019) [34], 2: RetinaFace (2020) [6], 3: HAMBox (2020) [41], 4: TinaFace (2020) [9], 5: SCRFD-2.5GF (2021) [29], 6: TinyYolov3 (2022) [42], 7:YOLOv7-tiny-Face (2022) [7], 8: YFaces-Tiny (2024) [43], 9: YOLOv8n-face (2023) [7], 10: YOLOv5s-Face (2021) [7], 11: ADYOLOv5-Face (ours).

4.4.2. Experiments on XD-Face

To comprehensively assess the ADYOLOv5-Face model's capability in detecting small target faces, we conducted evaluation experiments on the proposed XD-face dataset. YOLO-based face detection methods were chosen as comparison benchmarks, and YOLO series evaluation metrics (precision, recall, AP50 and AP50:5:95) were utilized to verify detection accuracy. The data presented in Table 5 were obtained by testing the trained models available on the official websites of these face detectors. As depicted in Table 5, the proposed ADYOLOv5-Face model exhibits superior performance in terms of precision and AP, whereas the YOLOv5s-Face model performs better in the recall metric. Figure 8 also provides a clearer illustration of the evaluation results on the AP metric.

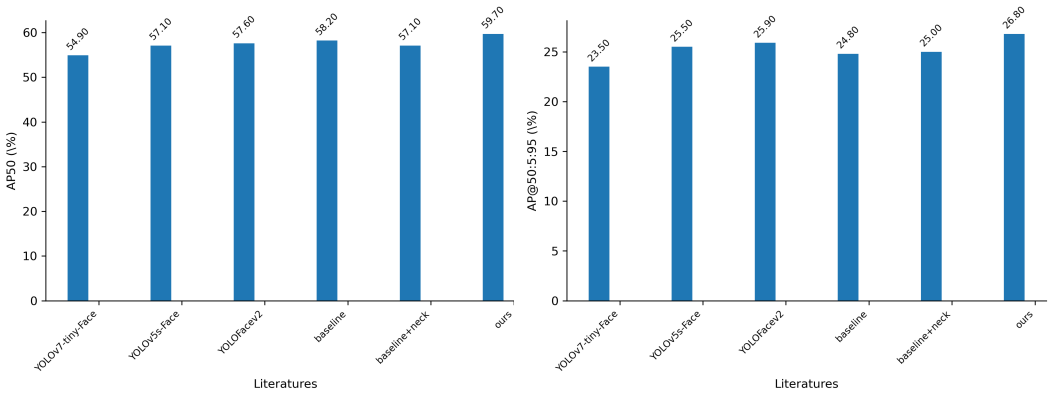


Figure 8. Detection results of the AP50 metric (**left**) and the AP@50:5:95 metric (**right**) on the XD-face dataset.

Given that the evaluation metric of AP50 (where IOU for the NMS is set to 0.5) is commonly used in object detection, we primarily analyze the performance of these face detectors based on AP50. Compared to the baseline method, the proposed model's performance increased by 1.5%, achieving a score of 59.7%. Furthermore, when compared to the other two YOLOv5-based face detectors (YOLOv5s-Face and YOLOfacev2), our proposed model demonstrates improvements of 2.6% and 2.1%, respectively. Additionally, our model outperforms the YOLOv7-tiny-based face detector by 4.8% in terms of the AP@50 metric.

Table 5. Comparison of our ADYOLOv5-Face with the YOLO-based face detectors on the XD-Face validation dataset under four metrics: precision, recall, AP50 and AP50:5:95.

Detector	Precision (%)	Recall (%)	AP50 (%)	AP@50:5:95 (%)
YOLOv7-tiny-Face	63.9	63.7	54.9	23.5
YOLOv5s-Face	46.8	74.6	57.1	25.5
YOLOFacev2	65.5	67.8	57.6	25.9
baseline	65.5	66.5	58.2	24.8
baseline+neck	65.7	66.7	57.1	25.0
ADYOLOv5-Face (ours)	66.5	70.2	59.7	26.8

In conclusion, when dealing with extremely small and occluded faces, our proposed model achieves superior detection accuracy compared to other state-of-the-art face detectors.

5. Conclusions

In this paper, we introduce ADYOLOv5-Face, which is based on the YOLOv5 object detector, and present a novel dataset (XD-face) for face detection tasks. We propose two variations of the YOLO structure, both of which demonstrate significant enhancements

to face detection performance. Through evaluation experiments conducted on the Wider Face and XD-face datasets, we demonstrate that ADYOLOv5-Face can effectively compete with other state-of-the-art face detectors while utilizing lightweight base models. Our results indicate that ADYOLOv5-Face either closely matches or even surpasses existing face detectors on both the Wider Face and XD-face benchmarks. These findings underscore the effectiveness of the advanced YOLOv5 architecture in achieving state-of-the-art performance, particularly in the detection of small faces.

The proposed method shows significant effectiveness in handling dense facial data and has potential applications in face counting, face recognition, and domain adaptation. In terms of face counting, it can more accurately identify and count the number of students in a classroom, contributing to the development of intelligent education. For face recognition, improving face detection accuracy serves as a fundamental step towards reliable facial identification. In domain adaptation, the algorithm can self-adjust to different classroom environments, enhancing recognition performance under varying lighting conditions and angles. Overall, the proposed algorithm provides a solid foundation for the application of facial recognition technologies in the education field, facilitating improvements in teaching efficiency and student management.

Author Contributions: Writing—original draft, L.L.; Visualization, G.W.; Funding acquisition, Q.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was jointly funded by the New Teacher Innovation Fund of Xidian University under Grant No. XJSJ23035, and the 2024 Higher Education Scientific Research Planning Project of the Chinese Association of Higher Education under Grant No. 241jpfhZD01.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

YOLO	You Only Look Once
CNN	Convolutional Neural Network
RFE	Receptive Field Enhancement module
FPN	Feature Pyramid Network
GD mechanism	Gather-and-Distribute mechanism
GHM	Gradient Harmonizing Mechanism
SWF	Slide Weight Function (SWF)
NWD	Normalized Wasserstein Distance
IoU	Intersection over Union
RPN	Region Proposal Network
FAM	feature alignment module
IFM	information fusion module
IIM	information injection module
Low-GD	low-stage Gather-and-Distribute mechanism (Low-GD)
AvgPool	average pooling operation
High-GD	high-stage gather-and-distribute mechanism
AP	Average Precision
PR	Precision Recall
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative

References

1. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
2. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
3. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part I 14; Springer: Cham, Switzerland, 2016; pp. 21–37.
4. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
5. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
6. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. Retinaface: Single-shot multi-level face localisation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5203–5212.
7. Qi, D.; Tan, W.; Yao, Q.; Liu, J. YOLO5Face: Why reinventing a face detector. In *Computer Vision—ECCV 2022 Workshops*; Springer: Cham, Switzerland, 2022; pp. 228–244.
8. Zhang, S.; Chi, C.; Lei, Z.; Li, S.Z. Refineface: Refinement neural network for high performance face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4008–4020. [CrossRef] [PubMed]
9. Zhu, Y.; Cai, H.; Zhang, S.; Wang, C.; Xiong, Y. Tinaface: Strong but simple baseline for face detection. *arXiv* **2020**, arXiv:2011.13183.
10. Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; Li, S.Z. Faceboxes: A CPU real-time face detector with high accuracy. In Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, 1–4 October 2017; pp. 1–9.
11. Ju, L.; Kittler, J.; Rana, M.A.; Yang, W.; Feng, Z. Keep an eye on faces: Robust face detection with heatmap-Assisted spatial attention and scale-Aware layer attention. *Pattern Recognit.* **2023**, *140*, 109553. [CrossRef]
12. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.
13. Chi, C.; Zhang, S.; Xing, J.; Lei, Z.; Li, S.Z.; Zou, X. Selective refinement network for high performance face detection. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8231–8238.
14. Chen, M.; Ren, X.; Yan, Z. Real-time indoor object detection based on deep learning and gradient harmonizing mechanism. In Proceedings of the 2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS), Liuzhou, China, 20–22 November 2020; pp. 772–777.
15. Cao, Y.; Chen, K.; Loy, C.C.; Lin, D. Prime sample attention in object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11583–11591.
16. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. Yolo-facev2: A scale and occlusion aware face detector. *Pattern Recognit.* **2024**, *155*, 110714. [CrossRef]
17. Jiang, H.; Learned-Miller, E. Face detection with the faster R-CNN. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 650–657.
18. Sun, X.; Wu, P.; Hoi, S.C. Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing* **2018**, *299*, 42–50. [CrossRef]
19. Zhu, C.; Zheng, Y.; Luu, K.; Savvides, M. CMS-RCNN: Contextual multi-scale region-based cnn for unconstrained face detection. In *Deep Learning for Biometrics*; Springer: Cham, Switzerland, 2017; pp. 57–79.
20. Khan, S.S.; Sengupta, D.; Ghosh, A.; Chaudhuri, A. MTCNN++: A CNN-based face detection algorithm inspired by MTCNN. *Vis. Comput.* **2024**, *40*, 899–917. [CrossRef]
21. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Huang, G.; Liu, Z.; Pleiss, G.; Van Der Maaten, L.; Weinberger, K.Q. Convolutional networks with dense connectivity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *44*, 8704–8716. [CrossRef] [PubMed]
24. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
26. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.

27. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
28. Liu, Y.; Tang, X. Bfbox: Searching face-appropriate backbone and feature pyramid network for face detector. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13568–13577.
29. Guo, J.; Deng, J.; Lattas, A.; Zafeiriou, S. Sample and computation redistribution for efficient face detection. *arXiv* **2021**, arXiv:2105.04714.
30. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
31. Zhang, J.; Wu, X.; Hoi, S.C.; Zhu, J. Feature agglomeration networks for single stage face detection. *Neurocomputing* **2020**, *380*, 180–189. [CrossRef]
32. Najibi, M.; Samangouei, P.; Chellappa, R.; Davis, L.S. SSH: Single Stage Headless Face Detector. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
33. Wang, W.; Wang, X.; Yang, W.; Liu, J. Unsupervised face detection in the dark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1250–1266. [CrossRef] [PubMed]
34. Li, J.; Wang, Y.; Wang, C.; Tai, Y.; Qian, J.; Yang, J.; Wang, C.; Li, J.; Huang, F. DSFD: Dual Shot Face Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
35. Zhang, Z.; Shen, W.; Qiao, S.; Wang, Y.; Wang, B.; Yuille, A. Robust face detection via learning small faces on hard images. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 1361–1370.
36. Fang, Z.; Ren, J.; Marshall, S.; Zhao, H.; Wang, Z.; Huang, K.; Xiao, B. Triple loss for hard face detection. *Neurocomputing* **2020**, *398*, 20–30. [CrossRef]
37. Wu, S.; Li, X.; Wang, X. IoU-aware single-stage object detector for accurate localization. *Image Vis. Comput.* **2020**, *97*, 103911. [CrossRef]
38. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.
39. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.M.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
40. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
41. Liu, Y.; Tang, X.; Han, J.; Liu, J.; Rui, D.; Wu, X. Hambox: Delving into mining high-quality anchors on face detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13043–13051.
42. Gao, J.; Yang, T. Face detection algorithm based on improved TinyYOLOv3 and attention mechanism. *Comput. Commun.* **2022**, *181*, 329–337. [CrossRef]
43. Sufian Chan, A.A.; Abdullah, M.; Mustam, S.M.; Poad, F.A.; Joret, A. Face Detection with YOLOv7: A Comparative Study of YOLO-Based Face Detection Models. In Proceedings of the 2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), Miri Sarawak, Malaysia, 17–19 January 2024; pp. 105–109.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.