Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3, 840-855
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i3.5356

GCSEM-YOLO small scale enhanced face detector based on YOLO

Xuwen Zheng[1], Zhiwei Zhou[2], Chonlatee Photong[3*]
[1,3]Faculty of Engineering Mahasarakham University, Maha Sarakham, 44150, Thailand; 2240053955@qq.com (X.Z.) chonlatee.p@msu.ac.th (C.P.).
[2]Hunan Mechanical & Electrical Polytechnic, Changsha, 410151, Hunan, China; 65010363010@msu.ac.th (C.P.).

Abstract: Face detection is a crucial aspect of computer vision, often challenged by factors such as varying scales, occlusions, and diverse facial features. In this study, we introduce GCSEM-YOLO, an innovative real-time face detection method built upon the YOLOv8 architecture. This approach incorporates a novel feature extraction module (GCSEM) alongside a specialized small-scale detection head, designed to capture pixel information across multiple levels and enhance the receptive field, thereby improving the accuracy of small face detection. To address the imbalance between easy and difficult samples, we employ an adaptive anchor box filtering algorithm coupled with the new WIoUv3 loss function. Experimental results demonstrate that GCSEM-YOLO achieves outstanding efficiency on the WIDER FACE validation datasets.

Keywords: *Face recognition, GCSEM-YOLO, Small scale, YOLO.*

1. Introduction

Face detection serves as a crucial foundation for enabling a variety of facial analysis tasks, particularly facial alignment Meher, et al. [1] attribute analysis Dey, et al. [2] recognition Du, et al. [3] and verification [4]. The performance of facial detectors has significantly increased with the recent explosion in deep convolutional neural networks. Several deep learning-based high-performance facial detection systems have been put forth. These algorithms often fall into two major categories. One category is cascaded neural nets. The face detection technique detects coarse to fine faces by using cascaded neural networks as classifiers and feature extractors. Although cascade detectors have achieved notable success, they also present certain limitations, including challenges in training and slower detection speeds. Another class of algorithms stems from improvements to generalized target detection algorithms. With the use of specialized architectural designs, task-specific models can improve performance by utilizing the more universal object properties that these detectors are made to capture. Prominent face detectors like YOLO [5]. Faster R-CNN Nogales, et al. [6] and RetinaNet exemplify this approach. In the research, inspired by YOLOv5 [7]. They present GCSEM-YOLO, a revolutionary face detector that reaches cutting-edge results in single-stage face detection.

Even though face detection performance has significantly improved thanks to deep convolutional networks, it is still difficult to distinguish faces with large scale differences in real-world situations. Unlike general-purpose object detection, face detection involves relatively small aspect ratio variations (typically between 1:1 and 1:1.5) but substantial scale variations, with face sizes ranging from a few pixels to thousands of pixels [8]. This continuous variation in scale complicates adaptation for discrete anchor-based algorithms, making it difficult to accurately detect faces across diverse scales. To tackle this issue, scholars have suggested various approaches.Du et al. proposed YOLO-Face Du, et al. [3] an enhanced face detector based on YOLOv3 that focuses on scale variation. The method involves using a more accurate regression loss function and creating anchor scales that are especially suited for human

History: Received: 6 January 2025; Revised: 25 February 2025; Accepted: 27 February 2025; Published: 11 March 2025
* Correspondence: chonlatee.p@msu.ac.th

faces. For example, Hossain et al. [9] optimized anchor sizes based on the effective receptive field, allowing the model to better match the natural variation in face sizes. Similarly, FaceBoxes Qi, et al. [10] introduced multi-scale anchors, which enhance the receptive field by assigning different anchor sizes to various layers, thereby improving the model's ability to detect faces at a wide range of scales. These adaptations help ensure that the network can effectively capture faces of varying sizes, enhancing detection accuracy and robustness in complex, real-world scenarios. Motivated by the idea of receptive fields, we created a GCSEM module that offers a greater variety of receptive fields, improving the model's capacity to identify faces at various scales. Numerous studies have aimed to advance face detection architectures, primarily divided into single-stage Zhang, et al. [11] and two-stage Wu, et al. [12] detectors.

Densely packed anchor points are deployed throughout all locations of numerous multi-scale feature maps with different sizes and aspect ratios by single-stage detectors, which are domain- and anchor-point-based. By utilizing the flattened anchor points, this method enables the model to effectively recognize faces in a single pass, providing a simplified and effective solution that is especially advantageous for real-time applications. Two-stage detectors, on the other hand, usually use a region proposal network to produce candidate regions, which are then refined in a second stage. This increases accuracy but frequently comes at the expense of more computing complexity.

The head network, the feature module, the multitask loss, and the backbone are the four main parts of this system. The feature extraction module employs a Feature Pyramid Network (FPN) Huang, et al. [13] to effectively aggregate high- and low-level features from the backbone network, enabling the model to capture fine-grained and contextually rich details across scales. Additionally, modules for refining the receptive field, such as Receptive Field Blocks (RFBs) Terven, et al. [14] are incorporated to enhance contextual information, especially for challenging face detections in complex environments.

The multitask loss function used in this model includes both binary classification and bounding box regression. The binary classification loss categorizes predefined anchors as either faces or backgrounds, while the bounding box regression loss enhances the predicted locations of detected faces, ensuring precise localization. Together, these components contribute to a more robust face detection system that performs well across varied face sizes and challenging visual contexts.

However, difficult samples and small-size face detection issues remain a problem for many detection methods. To address this problem, we designed a novel feature extraction module that improves small-size and difficult sample detection while minimizing the loss of detail information. However, faces may appear extremely small in certain situations, such as surveillance settings. Ignoring the identification of these targets could cause an intelligent surveillance system to malfunction. Tiny faces, particularly those smaller than 16x16 pixels, are usually detected by densely tiling a large number of tiny anchor points throughout the image in modern face identification techniques. Although this method increases recall, it exacerbates the issue of high category imbalance, a significant cause of excessive false positives [15]. The cumulative effect of these false positives during training significantly affects the model's updates, resulting in a notable sample imbalance that compromises the training process' effectiveness. This imbalance often causes the model to focus disproportionately on background regions, reducing its ability to effectively learn distinctive features for accurate face detection, especially for small faces in cluttered or complex scenes. To deal with the problem, Lin et al. [16] introduced Focal Loss, which dynamically assigns greater weight to difficult samples, reducing the impact of easily classified examples and thus helping to alleviate class imbalance. Correspondingly, the Gradient Harmonizing Mechanism (GHM) Meher, et al. [1] reduces the gradient contributions of simple samples, allowing the model to prioritize learning from harder examples. Another approach, Prime Sample Attention (PISA) Song, et al. [17] assigns weights to both positive and negative samples according to distinct criteria, further enhancing the model's focus on challenging samples. These strategies collectively improve training efficiency and detection accuracy by mitigating the effect of sample imbalance and reducing false positives.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

However, current hard sample mining techniques are difficult to put into reality because they frequently call for adjusting a lot of hyperparameters. Furthermore, accurately detecting faces with substantial variations in scale, pose, occlusion, expression, appearance, and illumination remains a significant challenge [18]. In addition, the use of large models and high computational costs can limit the applicability of these detectors in real-world scenarios, particularly those with stringent memory and latency constraints. To handle these problems, we propose a novel model called ATW, designed to alleviate the problem of unbalanced samples. ATW enables the model to adaptively learn the threshold parameters for positive and negative samples, reducing reliance on manual tuning. Additionally, we incorporate the wIoUv3 Tong, et al. [19] loss to address the issue of slight Intersection over Union (IoU) bias, which is especially problematic in small-scale face detection. These innovations improve the efficiency and accuracy of face detection while ensuring that the model can be deployed in resource-constrained environments.

All in all, our main efforts are:

1.A new feature extraction module which leverages SE mechanisms and spatial attention.

2.Introduced the wIoUv3 loss function to enhance bounding box regression and applied an adaptive anchor frame filtering algorithm to address sample imbalance, improving the model's performance on challenging detection tasks.

3.To enhance small item identification, a tiny detection head is added.

2. Related Work

For many years, computer vision research has focused heavily on facial detection. Handcrafted features, like Haar features, were the mainstay of early facial detection techniques Alrammahi and Jabbar [20] control point sets, and edge direction histograms. Face detection methods started using neural networks for automatic feature extraction and classification with the introduction of deep learning. For instance, CascadeCNN Venieris, et al. [21] presented a cascade structure made up of three meticulously crafted deep convolutional networks that made coarse-to-fine predictions about landmark and facial positions. Similarly, MTCNN Khan, et al. [22] employed a cascading architecture that enabled joint alignment of facial landmarks and detection of facial positions. The PCN Kaewkorn, et al. [23] on the other hand, used a direction prediction network to correct facial orientations, further improving detection accuracy. These early advancements laid the groundwork for modern, more robust face detection methods. Early deep learning-based facial detection techniques, however, had a number of shortcomings, for instance laborious training procedures, a vulnerability to local optima, sluggish detection speeds, and poor accuracy.

The sliding window paradigm, which applies classifiers to dense grids of images, has roots extending back several decades. The landmark study by Viola-Jones [24] introduced a cascade framework that effectively eliminated erroneous facial regions within image pyramids in real-time, leading to the widespread adoption of this scale-invariant face detection method. While the sliding window approach Aggarwal, et al. [25] on image pyramids had long been the dominant paradigm for detection, the introduction of feature pyramids Zhao, et al. [26] has rapidly shifted the focus towards sliding anchors on multi-scale feature maps. This transition allows for more efficient and accurate face detection by leveraging rich hierarchical features, enabling better handling of faces at various scales and improving overall detection performance.

Current facial detection methods build upon advancements from general object detection techniques and are typically categorized into two main classes: two-stage methods (e.g., Faster R-CNN) and single-stage methods (e.g., SSD and RetinaNet). Two-stage methods employ a "proposal and refinement" mechanism, providing higher localization accuracy [27]. Single-stage approaches, on the other hand, sample facial positions and scales densely, which frequently leads to a notable imbalance between positive and negative data during training. This imbalance arises because the majority of anchor points correspond to background regions, while only a small fraction corresponds to actual faces, leading to a dominance of easy negative samples. This issue can hinder the model's ability to learn effectively,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

particularly for detecting smaller or more difficult faces, and may require additional strategies, such as loss weighting or hard sample mining, to improve performance. To address this imbalance, sampling and weighting techniques have been widely adopted. Compared to two-stage methods, single-stage approaches are more efficient and typically offer higher recall rates, as they perform detection in a single pass, making them well-suited for real-time applications. But this effectiveness may come at the expense of decreased localization accuracy and possibly increased false positive rates. The dense sampling of anchor points in single-stage methods can result in more background regions being misclassified as faces, and the lack of a refinement step, common in two-stage methods, can lead to less precise bounding box predictions. Balancing these trade-offs remains a key challenge in optimizing single-stage face detection models.

The benefits of general object detection methods, like SSD, have had a major impact on recent developments in facial detection algorithms [28]. Faster R-CNN Han, et al. [29] and RetinaNet [30]. FaceBoxes Qi, et al. [10] a lightweight network built on the SSD architecture, was created by Zhang et al. This algorithm efficiently reduces the feature size by a factor of 32 through down sampling, while simultaneously utilizing multi-scale network modules to enrich feature dimensions in both width and depth. This method enables the model to capture a broader spectrum of spatial information and scale variations, improving its ability to detect faces at different sizes and under varying conditions. By balancing dimensionality reduction with enhanced feature representation, the algorithm achieves both computational efficiency and improved detection accuracy. The SRN Yang, et al. [31] is an improved version of the RefineDet and RetinaNet object identification algorithms that achieves great performance by adding multi-branch modules to increase the impact of the receptive field and introducing two-stage classification and regression procedures. Furthermore, by combining facial alignment and facial detection, MTCNet utilizes facial landmarks to improve detection performance.

Facial identification algorithms face significant hurdles due to significant fluctuations in facial scale within complex scenarios, which negatively impacts their performance. Multi-scale detection capabilities are largely driven by scale-invariant features, which allow the model to detect objects, such as faces, across a wide range of sizes. This has led to extensive research focused on improving the accuracy and efficiency of feature extraction. By enhancing feature representations that are robust to scale variations, researchers aim to create models that can effectively detect faces at both small and large scales while maintaining high precision and computational efficiency. These advancements often involve the use of multi-scale feature pyramids, adaptive sampling strategies, and more sophisticated network architectures that better capture scale-invariant patterns. Using dilated convolutions and lowering the number of down sampling layers can greatly improve detection performance for small item detection. Using more anchors, which offer useful prior knowledge, is another way to address this issue; using denser anchors in conjunction with suitable matching techniques can significantly raise the calibre of object proposals. Multi-scale training provides a simple yet efficient method to improve multi-scale object detection performance, and it is essential for building image pyramids and expanding sample diversity [32]. By training on images at various scales, this technique allows the model to learn features that are robust to scale variations, improving detection across different object sizes. However, as the receptive field expands and deeper semantic information is integrated, there is a risk of losing fine-grained spatial details. A natural solution to this challenge is to combine deep semantic features with shallow features Kavitha and RajivKannan [33] thereby retaining spatial information while benefiting from richer context. This fusion enables the model to maintain both high-level semantic understanding and precise spatial localization, improving detection accuracy in complex scenarios.

In densely populated scenes, occlusion poses a significant challenge by causing incomplete data regarding occluded faces. As certain facial regions become hidden or boundaries become ambiguous, the model may fail to detect these faces, leading to missed detections and low recall rates. Occlusion can obscure key facial features or cause overlapping objects to be misclassified, reducing the effectiveness of traditional detection methods. To address this, more advanced strategies, such as leveraging contextual information, multi-view detection, or incorporating occlusion-aware models, are needed to improve

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

detection accuracy in such complex scenarios. Contextual information is crucial for mitigating occlusion problems in facial detection, according to earlier research. SSH Zhang, et al. [11] utilizes simple convolutional layers to extend the windows around candidate regions, effectively incorporating contextual information to help detect faces that may be partially occluded. FAN Wang, et al. [34] introduces anchor attention, which prioritizes facial features, allowing the model to focus on key areas even when faces are partially hidden. PyramidBox Cao, et al. [35] enhances this approach by implementing a context-sensitive prediction module, replacing SSH's convolutional layers with DSSD's residual prediction module, further improving its ability to handle occlusions. Additionally, RetinaFace Hao, et al. [36] applies independent contextual modules across five feature pyramid layers, expanding the receptive field and strengthening the model's capacity to model rigid context. These methods have shown strong performance in mitigating occlusion issues, demonstrating that leveraging contextual information is an effective strategy for improving detection in occluded areas. This line of research offers promising potential and warrants further exploration to enhance robustness in challenging detection environments.

The total loss function is dominated by easy samples, which make up a sizable part of samples in the single-stage facial detection domain. Consequently, learning features from simple samples is frequently given priority in models, whereas the more difficult and complex examples are underrepresented throughout the training phase. Poor generalization may result from this, especially when there is occlusion, small faces, or odd postures. To tackle this problem, the Online Hard Example Mining (OHEM) approach integrates hard examples into the stochastic gradient descent (SGD) training process by choosing them according to their loss values. By focusing the model's attention on these difficult examples, OHEM helps ensure that the model learns to detect faces under more challenging conditions, therefore enhancing overall detection accuracy and robustness [37].

3. Method

3.1. GCSEM-YOLO Model Overview

The structure of GCSEM-YOLO is depicted in Figure 1.We have made many adjustments and enhancements to the YOLOv8n fundamental design, with a particular emphasis on the neck and backbone parts.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

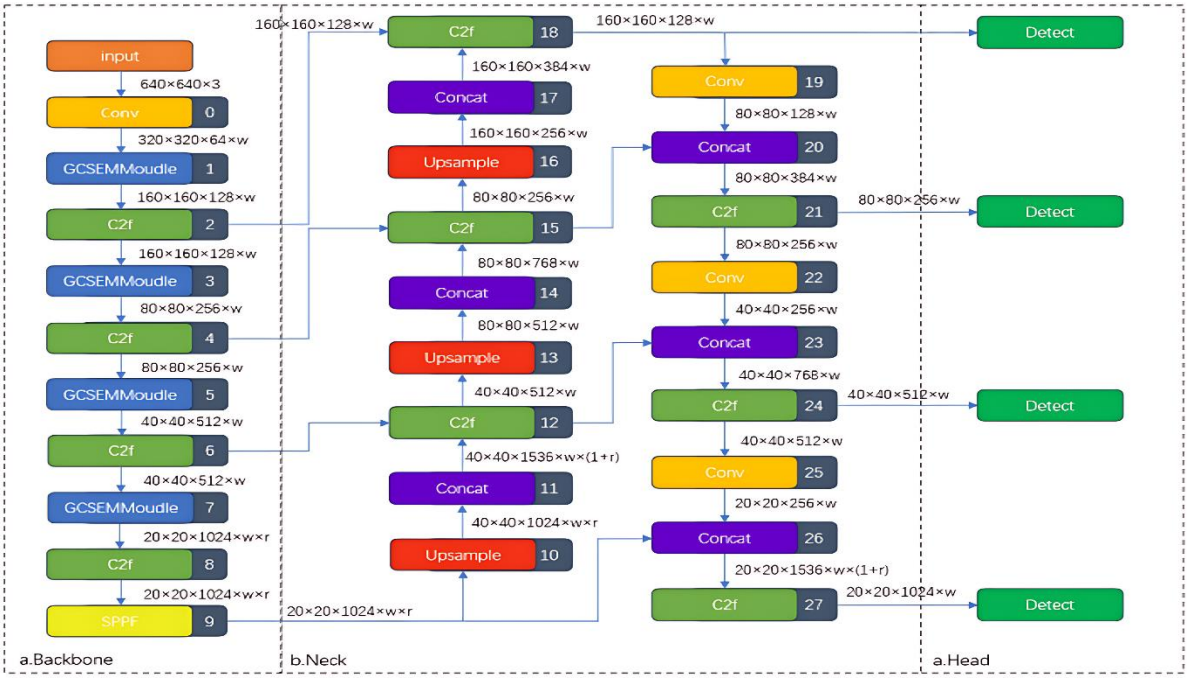

Figure 1.
YOLOv8's Network Structure.

The size of the feature map is represented by the parameters w (width) and r (ratio) in Figure 1. By adjusting the values of w and r to suit the requirements of various application scenarios, the model's size may be managed. Four feature maps with dimensions of 160 × 160, 80 × 80, 40 × 40, 20 × 20 are produced by the network.

3.2. Feature Extraction Module with Squeeze-and-Excitation (SE) and Group Convolutions

In our endeavour to bolster the feature extraction process of YOLOv8, we have engineered an innovative GCSEM Module that seamlessly integrates the Squeeze-and-Excitation (SE) mechanism and spatial attention, enhancing the network's ability to discern and emphasize pivotal features within the input data.

Group Convolution: In our module, we use group convolutions to balance feature representation with parameter economy. By dividing the input channels into groups, this method enables the network to concentrate on local groups of channels and acquire more generalized features. This method lessens the computational load related to managing big parameter sets while simultaneously enhancing the model's capacity to identify various and pertinent patterns across various spatial locations. The technique improves computing efficiency without compromising the model's capacity to identify intricate relationships in the data by reducing the number of learning parameters. As a result, the network architecture becomes more effective and scalable. Mathematically, the feature extraction process within our module can be articulated as:

$$X = \text{conv1}(\text{conv2}(x)) \tag{1}$$

where conv1 denotes a group convolution operation, and conv2 represents a standard convolution layer tasked with downsampling the feature map, the input data is x

Squeeze-and-Excitation (SE) Mechanism: The SE module Cai, et al. [38] plays a pivotal role in recalibrating channel-wise feature responses by adaptively scaling them based on their global

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

importance. This recalibration improves feature representation and discrimination by allowing the model to preferentially focus on the most important spatial information. By emphasizing relevant features and suppressing less informative ones, the SE module not only refines the model's performance in detecting key objects but also boosts its robustness against varying contextual changes, such as illumination or pose variations, making it more effective in dynamic and complex environments. After the first step of convolution, the feature map X becomes the feature map U:

$$\mathrm{U_c} = \sum_{\mathrm{s=1}}^{\mathrm{C'}} \mathrm{V_c^s} * \mathrm{X^s} \tag{2}$$

To overcome the challenge of exploiting channel dependency, the feature map containing global information is compressed directly into a feature vector, and the global spatial information is selected to be compressed into a single channel descriptor, i.e., the global average pooling of channels is employed. Because all of the channel features of the feature maps are condensed into a single numerical value, the issue of channel reliance is lessened and contextual information is included in the resulting channel-level statistics. The following is the definition:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{3}$$

Spatial Attention: We take a two-pronged approach to spatial attention, extracting the most important spatial cues for each channel by combining adaptive max pooling with adaptive average pooling. In particular, both average pooling and maximum pooling are used to process the input feature maps initially. To create a weight vector, these two pooling results are concatenated and then run through a convolutional layer and a Sigmoid function. The weighted feature maps are then created by multiplying this weight vector by the input feature maps. This dual strategy ensures a comprehensive capture of spatial features by leveraging both average and extreme responses, thereby equipping the model with enhanced sensitivity to the geometric configuration of the input data and improving its ability to focus on relevant spatial information. Max Pooling definition is as follows:

$$M(X) = \max_{i,j} X(i,j) \tag{4}$$

Within the channel dimension, the outcomes of maximum pooling and average pooling are then combined:

$$C\big(z_c, M(x)\big) = [z_c; M(X)] \tag{5}$$

After a layer of convolution and sigmoid activation:

$$K(C) = W \cdot C + b \tag{6}$$

$$S(K) = \frac{1}{1 + e^{-K}} \tag{7}$$

The weighted feature map is created by comparing the weight vector produced by the Sigmoid function with the initial input feature map:

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

$$W(x) = S \odot x \tag{8}$$

The GCSEM Module's capacity to lessen the loss of spatial information while down sampling is a key benefit. Preserving sufficient spatial information in large-scale feature maps is crucial for effective target detection, especially in face recognition scenarios. Although down sampling improves overall semantic understanding and extracts high-level semantic information, it frequently results in the loss of geographical information. In order to solve this issue, the GCSEM Module maintains spatial information during down sampling. By explicitly modelling the interdependencies between the convolved feature channels, broad local receptive field features are extracted from the feature map. This approach captures detailed and localized patterns in the input data, which are crucial for accurate object detection. Subsequently, spatial information extraction leverages these enhanced receptive field features, allowing the model to effectively integrate both local and contextual information. This enriched representation helps the model focus on important spatial cues and improves its ability to handle complex visual tasks, such as detecting faces at various scales and under different conditions.

Even after the GCSEM module is executed and the feature map is restored to its original size, the expanded intermediate receptive field feature maps allow for a full extraction of the spatial information of small objects. The GCSEM module outperforms traditional down sampling convolutions by significantly reducing the loss of spatial information. It enhances the fidelity of spatial information within the feature maps, ensuring that fine-grained details, particularly for small objects, are preserved. This improvement helps the model maintain accurate spatial representation, which is crucial for detecting small and challenging objects in complex visual scenes.

3.3. WIoUv3 Loss Function and Adaptive Filtering

We present the WIoUv3 loss function to improve the bounding box regression component of YOLOv8, a novel approach designed to overcome the limitations of traditional IoU-based loss functions, especially when addressing minor inaccuracies in bounding box predictions. WIoUv3 integrates a dynamic, non-monotonic focusing mechanism in contrast to traditional loss functions that depend on a static focusing mechanism. In addition to considering the aspect ratio, centroid distance, and overlap area, it dynamically modifies the focus according to the prediction quality. The model's capacity to manage small-scale faces or objects with minute localization mistakes is enhanced by this adaptive method, leading to more accurate bounding box predictions and better overall performance in challenging detection scenarios. WIoU assesses the anchor box's quality using a sensible gradient gain allocation technique. Three iterations of WIoU were proposed by Tong et al. [19] . Attention-based projected box loss was used in the design of WIoUv1, and focusing coefficients were introduced in WIoUv2 and v3. WIoUv3 innovatively incorporates adaptive weights, denoted by *r*, which are dynamically adjusted based on the delta and alpha parameters. These parameters are sensitive to the scale of the bounding boxes, ensuring a balanced error distribution across predictions of varying sizes. The WIoUv3 loss function is articulated as follows:

$$L_{WIoUv3} = r \times L_{WIoUv1} \tag{9}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{10}$$

$$\beta = \frac{L^{*}_{IoU}}{\bar{L}_{IoU}} \in [0, +\infty) \tag{11}$$

Where is calculated based on the IoU target values and mean IoU, ensuring smooth adjustments to loss gradients. and are hyperparameters in the equation that can be changed to suit various models.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

Building upon WIoUv3, we introduce an adaptive filter that effectively filters out most easily distinguishable negative samples, thereby reducing the number of samples the classifier must process. This operation addresses a key challenge: when detecting small-scale faces, anchors are densely sampled from the shallow feature maps of the feature pyramid. While this leads to good recall for small-scale faces, it also adds a large number of false positives, which exacerbates the imbalance between positive and negative samples. Our adaptive filter greatly lessens the model's computational load by automatically eliminating the majority of easily separable negative samples.

The ATW module takes the coordinates and scores of anchor boxes as inputs, using a percentile-based approach to determine the filtering threshold. The threshold is dynamically calculated based on the score distribution of the anchors. For instance, if the percentile is set to 95, the threshold corresponds to the 95th percentile of the score distribution. This dynamic adjustment allows the threshold to automatically adapt to the score distribution, ensuring that it is optimized for different datasets and tasks, ultimately improving the efficiency and effectiveness of the model.

3.4 Small-Scale Detection Head

To tackle the detection of small-scale targets while considering the platform's resource limitations, this paper strategically integrates the efficient GCSEM Module to design the feature fusion network. An additional detection scale is introduced to complement YOLOv8's original three, enhancing the fusion of shallow features with more comprehensive positional information. The GCSEM Module is strategically inserted as a feature processing block the backbone network. As can be seen in Figure 2, the incorporation of the GCSEM Module serves to mitigate the resource demands associated with multi-scale feature fusion. The resultant model, now capable of four-scale detection, witnesses a marked enhancement in detection performance, particularly for small-scale targets, thus pushing the boundaries of what is achievable within the constraints of limited computational resources.

Figure 2.
The Terms large, Medium, Tiny, and xsmall are used to Indicate an Object's Size.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

4. Experiment

4.1. Wider Face Datasets

With 32,203 photos and 393,703 tagged faces—158,989 of which are in the training set and 39,496 in the validation set—the WIDER FACE datasets serve as a standard for face detection. There are three detection complexity levels in each subset: Easy, Medium, and Hard. As can be seen in Figure 3, the scale, placement, lighting, expression, and obstruction of these faces differ significantly.

Figure 3.
Examples of the WIDER FACE Dataset.

These public datasets WIDER is the primary source of the photographs chosen for WIDER FACE. The 61 WIDER event categories were chosen by the creators, who are from the Chinese University of Hong Kong. For each category, 40%, 10%, and 50% were chosen at random to serve as the training, validation, and testing sets.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

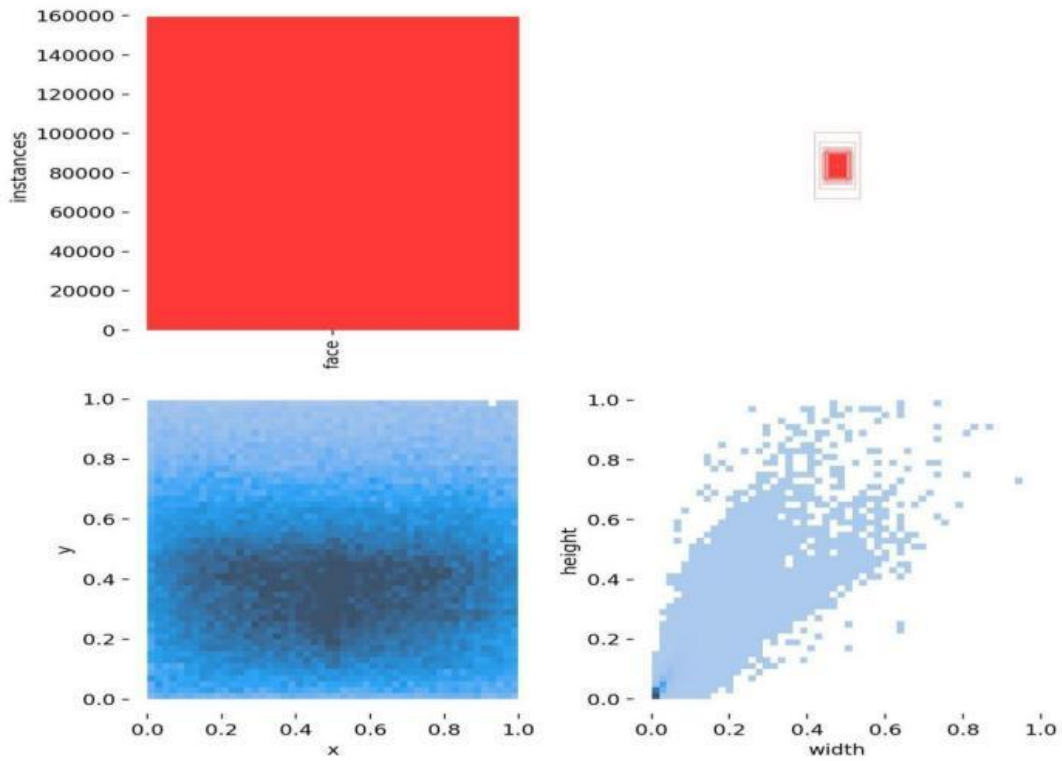

Figure 4.
Manually Labelled Object Details in the WIDER FACE Dataset

Python 3.8, PyTorch 2.1, and Cuda 11.8 were used in the desktop computing environment for the trials, which were conducted on a Windows 11 operating system. An NVIDIA 3090 graphics card was used as the hardware for these tests. The code for implementing the neural network was modified from Ultralytics version 8.0.202. The input images were resized to 640 × 640 pixels, and training was carried out for 300 epochs. Optimization was performed using the stochastic gradient descent (SGD) algorithm, starting with an initial learning rate of 0.01, which was progressively reduced to 0.0001. To ensure a fair and unbiased comparison among models, each was trained from scratch without employing pre-trained weights. This configuration made it possible to assess the model's performance only using its architecture and training.

4.2. Indicators of Experimental Evaluation

The suggested model's detection capabilities and model size were assessed. The accuracy of each object category was assessed using the precision (P), recall (R), average precision (AP), and mean average precision (mAP). Additionally, the giga floating-point operations per second (GFLOPs) metric was used to evaluate the networks' computational complexity. Frames per second (FPS) was used to gauge the model's real-time processing performance, and the number of parameters dictated the model's size.

Precision is defined as follows: Precision is the ratio of all detected samples to the number of positive samples predicted by the model:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

Precision is calculated as follows: Precision is the ratio of all detected samples to the number of positive samples identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

The area under the precision-recall curve is the mean precision (AP), which is calculated as follows:

$$\text{AP} = \int_0^1 \text{Precision}(\text{Recall}) d(\text{Recall}) \tag{14}$$

The model's detection performance across all categories is measured by mean average precision(mAP), which is the outcome of the weighted average of AP values for each sample category. The formula is displayed below:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{15}$$

The value having a category index value of is indicated in the equation above. indicates how many sample categories there are in the training dataset. When the detection model IOU is set to 0.5, the average accuracy is indicated by mAP0.5: when the detection model IOU is set to 0.5 to 0.95 (taken at 0.5 intervals), the average accuracy is indicated by Map 0.5:0.95.

5. Yolov8 Comparative Analysis

Table 1and Table 2 as well as Figure 5 display the GCSEM -YOLO and YOLOv8 models' sizes and performance on the Wider Face dataset.

Table 1.
A Comparison Between YOLOv8 and GCSEM-YOLO.

Network	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	FPS
YOLOv8n	84.7	58.0	66.7	36.6	162
GCSEM-YOLO-n	89.3	69.9	78.8	45.1	142
YOLOv8s	85.2	58.6	69.5	40.9	148
GCSEM -YOLO-s	92.2	72.9	82.5	47.1	117
YOLOv8m	87.6	60.9	72.3	43.6	117
GCSEM -YOLO-m	95.3	76.8	88.1	49.5	99
YOLOv8l	88.9	65.1	76.9	49.2	94
GCSEM -YOLO-l	98.4	82.9	95.1	55.7	88

In terms of detection accuracy at various scales, the results in Table 1 show that GCSEM-YOLO performs better than the conventional YOLOv8 model. In particular, GCSEM-YOLO-n outperforms YOLOv8n by 12.1%, achieving a mAP50 score of 78.8%. Furthermore, at all scales, GCSEM-YOLO attains better accuracies based on an FPS that is just marginally less than YOLOv8. According to these findings, GCSEM-YOLO reduces computational costs and increases detection accuracy while using a smaller model size.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

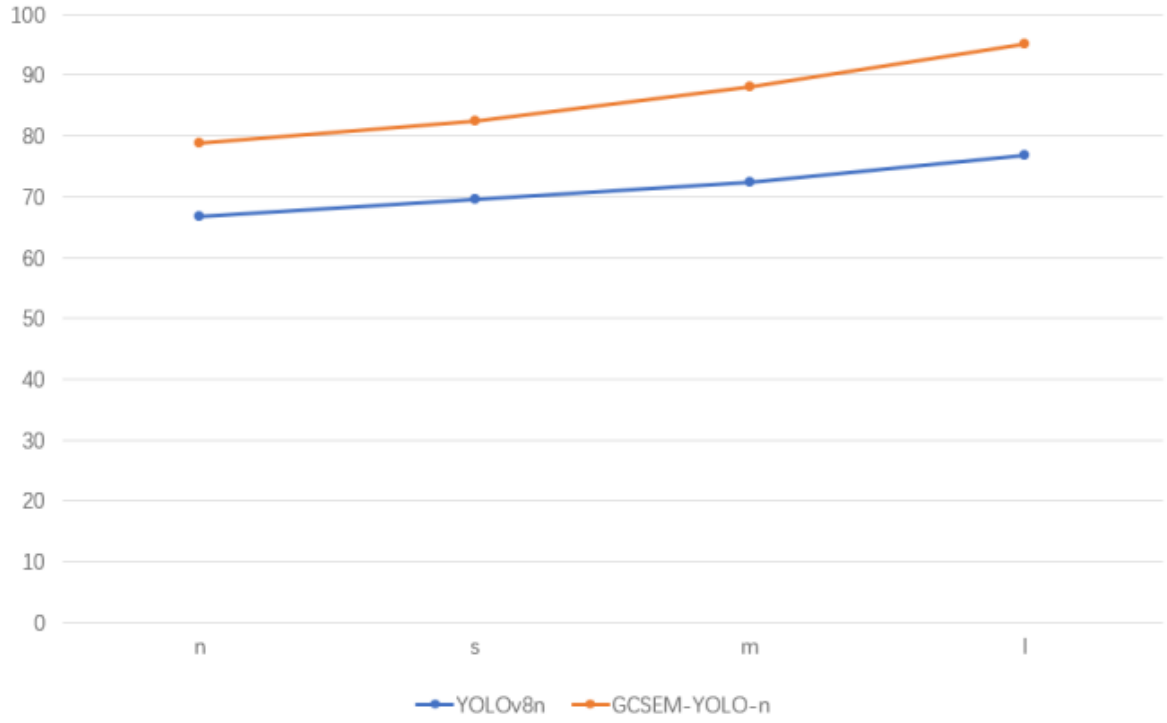

Figure 5.
Detection Accuracy Versus Scale.

Regardless of the scale, Figure 5 shows that the suggested detection model continuously performs better than the baseline model. Subsequent study shows that the suggested model's detection performance differs dramatically from the baseline model (in this case, YOLOv8) as the model size grows. According to the mAP50 indicator, GCSEM-YOLO-n performs 12.1% better than YOLOv8 in the "n"-scale setup. Nevertheless, the mAP50 indication is 18.2% larger when the scale is raised to "l," greatly increasing the difference between the two models. This occurrence suggests that as the number of model parameters and computational complexity rise, GCSEM-YOLO's detection accuracy improves more quickly. Therefore, when implemented on a hardware platform with adequate computational capabilities, the suggested model should offer high object-detection accuracy, particularly in large-scale settings.

Table 2.
Comparing The Detection of Faces Of Varying Sizes Using Yolov8.
Network AP-Small (%) AP-Medium (%) AP-Large (%)

Network	AP-easy (%)	AP-Medium (%)	AP-hard (%)
YOLOv8n	75.6	75.1	61.3
GCSEM-YOLO-n	91.5	89.9	79.3
YOLOv8s	78.4	78.2	61.9
GCSEM -YOLO-s	92.0	90.1	81.1
YOLOv8m	79.1	79.0	62.1
GCSEM -YOLO-m	92.9	90.8	83.5
YOLOv8l	82.7	82.2	65.8
GCSEM -YOLO-l	94.3	93.0	84.7

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

Table 2 shows that GCSEM-YOLO outperform YOLOv8 in dealing with large-size objects. In addition, GCSEM-YOLO also outperforms YOLOv8 in detecting hard targets, especially in the l-size model, where GCSEM-YOLO-l outperforms YOLOv8l by 18%-21.4%. According to these findings, GCSEM-YOLO offers notable benefits when it comes to identifying tiny and medium-sized faces, particularly small faces.

6. Ablation Experiment

This study carried out a number of compared experiments to evaluate the effects of including the GCSEM module in the model and to make ablation trials easier in the future. At various backbone network tiers, the GCSEM module was utilized in place of the Conv module, with YOLOv8n serving as the baseline model. Table 3 displays the experimental outcomes, with the "+" sign denoting modifications to the baseline model. According to the findings, GCIM-YOLO-N's mAP50 was 10 percentage points greater than the original YOLOv8n and 3 percentage points higher than YOLOv8n. The effectiveness of the GCSEM module was validated by ablation tests on the WIDER FACE dataset, which showed that the suggested enhancement significantly increased the accuracy of face detection.

A: Instead of using the original IoU loss function, the WIoUv3 loss function is used. Add ATW module before computing wiouv3 loss.

B: Add the *x*small-detection head.

C: The GCSEM module is utilized as the feature extraction module in place of the Conv module.

Table 3.
Ablation Study Results on the wider face dataset.

Network	Precision/%	Recall/%	mAP0.5/%	mAP0.5:0.95/%
YOLOv8s	84.7	58.0	66.7	36.6
+A	85.9	60.8	71.2	39.0
+A+B	86.4	62.1	72.9	41.1
+A+B+C	89.3	69.5	78.8	45.1

As observed from the experimental results in Table 3, each improvement strategy contributes to varying degrees of enhancement in detection performance when applied to the baseline model. The introduction of WIoUv3 in the prediction box regression loss, combined with the ATW module, improves the model's localization ability through a smarter sample allocation strategy, leading to a 4.5% increase in mAP50. Adding the *x*small detection head to the detection head architecture improved the mAP50 by 1.7%. An effective attention mechanism that improves concentration on important information in the feature map is integrated into the GCSEM module, which takes the place of the original Conv module in the backbone. As a result, mAP50 significantly improves by 5.9%.

Since there are a lot of little targets in the dataset, the GCSEM module's fusion of shallow feature information successfully lowers the rate at which small targets leak out. All things considered, our suggested feature fusion network shows promise in significantly enhancing the model's detection performance, especially for small-scale faces.

7. Conclusions

In order to tackle the problem of small target recognition in face detection tasks, we suggest the GCSEM-YOLO model, which is based on YOLOv8. The model successfully addresses typical face detection problems. The new GCSEM neck structure, which makes multi-scale feature fusion easier, is a significant algorithmic advance. By achieving a balanced integration of spatial and semantic information in the feature map, this structure enhances the model's capacity to identify faces at various scales. GCSEM decreases the sparsity of spatial information brought on by down sampling and improves feature extraction performance as compared to the conventional convolutional layer. Furthermore, a key factor in enhancing detection performance for small faces is the *x*small detection head, which is

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 840-855, 2024
DOI: 10.55214/25768484.v9i3.5356

especially made for small target face detection. The computational complexity of the model is further decreased, increasing its efficiency, through the application of WIoUv3 loss and a novel anchor frame filtering approach. GCSEM-YOLO performs better than YOLOv8 in terms of detection accuracy on practically all scales, according to experimental results on the WIDER FACE dataset. These findings indicate how well the GCSEM-YOLO model works to improve tiny target detection for facial photos.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] J. Meher, H. Allende-Cid, and T. E. Nordling, "A survey and classification of face alignment methods based on face models," *arXiv preprint arXiv:2311.03082*, 2023.

[2] A. Dey, T. Senapati, M. Pal, and G. Chen, "Pythagorean fuzzy soft RMS approach to decision making and medical diagnosis," *Afrika Matematika*, vol. 33, no. 4, p. 97, 2022. https://doi.org/10.1007/s13370-022-01031-7

[3] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1-42, 2022. https://doi.org/10.1145/3507902

[4] J. You, Y. Sun, M. Guo, C. Feng, and J. Ran, "LVFace: Large vision model for face recogniton," *arXiv preprint arXiv:2501.13420*, 2025.

[5] N. Algaraawi, T. Morris, and T. F. Cootes, "Facial feature point detection under large range of face deformations," *Journal of Visual Communication and Image Representation*, vol. 103, p. 104264, 2024. https://doi.org/10.1016/j.jvcir.2024.104264

[6] A. Nogales, M. Rodríguez-Aragón, and Á. J. García-Tejedor, "A systematic review of the application of deep learning techniques in the physiotherapeutic therapy of musculoskeletal pathologies," *Computers in Biology and Medicine*, vol. 172, p. 108082, 2024. https://doi.org/10.1016/j.compbiomed.2024.108082

[7] M. Jani, J. Fayyad, Y. Al-Younes, and H. Najjaran, "Model compression methods for YOLOv5: A review," *arXiv preprint arXiv:2307.11904*, 2023.

[8] B. B. Topal, "Reproducibility Report RetinaFace: Single-shot Multi-level Face Localization in the Wild." https://doi.org/10.1109/cvpr42600.2020.00525

[9] M. I. Hossain and H. Kabir, "An efficient way to recognize faces using mean embeddings," in *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2021: IEEE, pp. 1-10, doi: https://doi.org/10.1109/icaect49130.2021.9392401.

[10] S. Qi, J. Zhang, Z. Bao, and D. Li, "Making FaceBoxes better and faster by reparameterization," in *Third International Conference on Control and Intelligent Robotics (ICCIR 2023)*, 2023, vol. 12940: SPIE, pp. 726-729.

[11] J. Zhang, C. Hou, X. Yang, X. Yang, W. Yang, and H. Cui, "Advancing face detection efficiency: Utilizing classification networks for lowering false positive incidences," *Array*, vol. 22, p. 100347, 2024. https://doi.org/10.1016/j.array.2024.100347

[12] Q. Wu, X. Wang, N. Li, S. Fong, L. Zhang, and J. Yang, "Real-time face and facial landmark joint detection based on end-to-end deep network," *IEEE Transactions on Instrumentation and Measurement*, 2025. https://doi.org/10.1109/tim.2025.3541698

[13] S. Huang, Z. Lu, R. Cheng, and C. He, "FaPN: Feature-aligned pyramid network for dense image prediction," in *Proceedings of the IEEE/CVF International Conference on Computer vision*, 2021, pp. 864-873, doi: https://doi.org/10.1109/iccv48922.2021.00090.

[14] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680-1716, 2023. https://doi.org/10.3390/make5040083

[15] Z. Wang, X. Xie, J. Yang, and G. Shi, "Soft focal loss: Evaluating sample quality for dense object detection," *Neurocomputing*, vol. 480, pp. 271-280, 2022. https://doi.org/10.1016/j.neucom.2021.12.102

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980-2988, doi: https://doi.org/10.1109/iccv.2017.324.

[17] P. Song, P. Li, L. Dai, T. Wang, and Z. Chen, "Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection," *Neurocomputing*, vol. 530, pp. 150-164, 2023.

[18] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *International Journal of Computer Vision*, vol. 131, no. 5, pp. 1141-1162, 2023. https://doi.org/10.1007/s11263-022-01739-w

[19] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," *arXiv preprint arXiv:2301.10051*, 2023. https://doi.org/10.1016/j.patcog.2022.109256

[20] H. K. Alrammahi and K. A. Jabbar, "Automatic facial expression detection using genetic algorithm with svm and haar wavelet features," in *International Conference on Machine Learning, IoT and Big Data*, 2024, pp. 329-348, doi: https://doi.org/10.1007/978-981-97-8160-7_26.

[21] S. I. Venieris, J. Fernandez-Marques, and N. D. Lane, "Mitigating memory wall effects in CNN engines with on-the-fly weights generation," *ACM Transactions on Design Automation of Electronic Systems*, vol. 28, no. 6, pp. 1-31, 2023. https://doi.org/10.1145/3611673

[22] S. S. Khan, D. Sengupta, A. Ghosh, and A. Chaudhuri, "MTCNN++: A CNN-based face detection algorithm inspired by MTCNN," *The Visual Computer*, vol. 40, no. 2, pp. 899-917, 2024. https://doi.org/10.1007/s00371-023-02822-0

[23] H. Kaewkorn, L. Zhou, and W. Li, "RP-Net: A robust polar transformation network for rotation-invariant face detection," *Pattern Recognition*, vol. 158, p. 111044, 2025. https://doi.org/10.1016/j.patcog.2024.111044

[24] D. M. Abdulhussien and L. J. Saud, "An evaluation study of face detection by viola-jones algorithm," *International Journal of health Sciences*, vol. 6, no. S8, pp. 4174-4182, 2022. https://doi.org/10.53730/ijhs.v6ns8.13127

[25] A. Aggarwal, V. Kumar, and R. Gupta, "Object detection based approaches in image classification: A brief overview," in *2023 IEEE Guwahati Subsection Conference (GCON)*, 2023: IEEE, pp. 1-6, doi: https://doi.org/10.1109/gcon58516.2023.10183609.

[26] G. Zhao, W. Ge, and Y. Yu, "GraphFPN: Graph feature pyramid network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2763-2772, doi: https://doi.org/10.1109/iccv48922.2021.00276.

[27] A. M. Ridha, M. J. Mohammed, H. A. Saber, M. H. Chyad, and M. H. Abdulameer, "Advanced deep learning approaches for early detection and localization of ocular diseases," *Edelweiss Applied Science and Technology*, vol. 8, no. 6, pp. 3708-3721, 2024. https://doi.org/10.55214/25768484.v8i6.2813

[28] F. Yang, L. Huang, X. Tan, and Y. Yuan, "FasterNet-SSD: A small object detection method based on SSD model," *Signal, Image and Video Processing*, vol. 18, no. 1, pp. 173-180, 2024. https://doi.org/10.1007/s11760-023-02726-5

[29] G. Han, S. Huang, J. Ma, Y. He, and S.-F. Chang, "Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 1, pp. 780-789, doi: https://doi.org/10.1609/aaai.v36i1.19959.

[30] Y. Du, H. Chen, Y. Fu, J. Zhu, and H. Zeng, "AFF-Net: A strip steel surface defect detection network via adaptive focusing features," *IEEE Transactions on Instrumentation and Measurement*, 2024. https://doi.org/10.1109/tim.2024.3398131

[31] Q. Yang, X. Shixiang, L. Yue, and J. Jianbin, "Skeleton detection based on local short-connection unidirectional fusion networks," *Journal of University of Chinese Academy of Sciences*, vol. 40, no. 2, p. 250, 2023. https://doi.org/10.1016/j.neunet.2023.07.051

[32] Y.-l. Li, Y. Feng, M.-l. Zhou, X.-c. Xiong, Y.-h. Wang, and B.-h. Qiang, "DMA-YOLO: Multi-scale object detection method with attention mechanism for aerial images," *The Visual Computer*, vol. 40, no. 6, pp. 4505-4518, 2024. https://doi.org/10.1007/s00371-023-03095-3

[33] M. Kavitha and A. RajivKannan, "Hybrid convolutional neural network and long short-term memory approach for facial expression recognition," *Intelligent Automation & Soft Computing*, vol. 35, no. 1, 2023. https://doi.org/10.32604/iasc.2023.025437

[34] G. Wang, J. Li, Z. Wu, J. Xu, J. Shen, and W. Yang, "Efficientface: An efficient deep network with feature enhancement for accurate face detection," *Multimedia Systems*, vol. 29, no. 5, pp. 2825-2839, 2023. https://doi.org/10.1007/s00530-023-01134-6

[35] Z. Cao, Y. Qin, Y. Li, Z. Xie, J. Guo, and L. Jia, "Face detection for rail transit passengers based on single shot detector and active learning," *Multimedia Tools and Applications*, vol. 81, no. 29, pp. 42433-42456, 2022. https://doi.org/10.1007/s11042-022-13491-x

[36] J. Hao *et al.*, "Sheep face detection based on an improved retinaface algorithm," *Animals*, vol. 13, no. 15, p. 2458, 2023. https://doi.org/10.3390/ani13152458

[37] S. Wang, "Measurement of the robustness and sustainability of China's economic development based on the shock of the COVID-19 epidemic," *Edelweiss Applied Science and Technology*, vol. 9, no. 1, pp. 173-197, 2025. https://doi.org/10.55214/25768484.v9i1.4022

[38] L. Cai, H. Li, W. Dong, and H. Fang, "Micro-expression recognition using 3D densenet fused squeeze-and-excitation networks," *Applied Soft Computing*, vol. 119, p. 108594, 2022. https://doi.org/10.1016/j.asoc.2022.108594