

GFSM-DETR: A Gated Frequency-Spatial Modulation Detector for Real-Time Small Face Detection

Haomin Li, Zhen Song, Yanxing Huang and Beilei Wang^(✉)

Software College of Northeastern University, Hunnan District Shenyang, 110169, China

wangbl@swc.neu.edu.cn

Abstract. Real-time and accurate dense small-scale face detection constitutes a critical technology in computer vision. However, in complex environments, this task presents significant challenges, including high target density, minute scales, frequent mutual occlusions, and background noise interference. These factors often cause models to lose critical fine-grained features, leading to severe missed detections. To address these challenges, this paper proposes a high-accuracy real-time detector (GFSM-DETR) specifically optimized for dense small-scale face detection. First, the GFSM-Fusion module is designed to utilize synergistic representation learning across the spatial and frequency domains, effectively extracting and amplifying high-frequency features of small targets while suppressing background noise. Furthermore, the ECA-CATM module is constructed to facilitate noise-resistant global context modeling through multi-stage information interactions within the spatial and channel domains combined with an additive attention mechanism. In addition, the proposed Focal-PIoUv2 loss function utilizes a dual-constraint mechanism to effectively filter gradient interference generated by low-quality predicted boxes. On the SCUT-Head and Brainwash datasets, the proposed model significantly outperforms the baseline RT-DETR and a series of mainstream detectors, achieving mAP_{50} scores of 95.2% and 95.7%, respectively. Compared to leading mainstream detectors including YOLOv11m and YOLO26m, our model realizes mAP_{50} gains of 2.5% and 2.0%, along with Recall improvements of 3.9% and 2.0% across the two datasets.

Keywords: Small face detection, Frequency-Spatial Modulation, RT-DETR, Global Context Modeling

1 Introduction

Face detection has emerged as a cornerstone technology in various real-world applications, encompassing biometric authentication, human-computer interaction (HCI), advanced video surveillance, and intelligent public safety infrastructures [1,2]. With the increasing complexity of urban environments, achieving real-time and highly accurate dense small face detection is of particularly significant practical importance.

Current methodologies are mainly classified into CNN-based YOLO-series models [3,4,5,6] that require Non-Maximum Suppression (NMS), and Transformer-based,

fully end-to-end DETR-series models [7]. Despite the high accuracy of mainstream DETR variants such as DINO [8], Deformable DETR [9] and Co-DETR[10], their substantial computational overhead and high inference latency hinder their deployment in real-time applications. As the first real-time DETR model, RT-DETR [11] tackles the slow inference problem of conventional DETR models through its efficient hybrid encoder. Meanwhile, RT-DETR preserves the NMS-free paradigm, facilitating accurate identification of overlapping small targets. Consequently, it demonstrates superior performance potential over the YOLO series in detecting small faces in complex scenes.

Nevertheless, despite the remarkable advancements of existing models, dense small face detection in complex environments remains highly challenging. In such scenarios, targets are densely distributed and mutually occluded, with extremely small scales for both the faces and their key features [2]. In addition, large-scale complex backgrounds—characterized by severe illumination changes, motion blur, and cluttered textures—further obscure the limited critical information of small targets. The confluence of these factors makes it difficult for models to accurately distinguish small faces from the background or other overlapping faces, leading to a high rate of missed detection. Therefore, our GFSM-DETR emphasizes suppressing background interference and improving the extraction and enhancement of essential local features of small face targets.

To address the aforementioned challenges, we propose GFSM-DETR specifically optimized for dense small face detection. Our contributions are summarized as follows:

1. We propose a novel **Gated Frequency-Spatial Modulation Fusion (GFSM-Fusion)** module supported by our proposed **Gated Frequency-Spatial Convolution (GFSCConv)**. By performing synergistic representation learning across the spatial and frequency domains, it effectively extracts and amplifies the fine-grained high-frequency texture features of small targets. This design mitigates feature attenuation in the multi-scale fusion stage and improves the recall of small targets.
2. By integrating **Efficient Channel Attention (ECA)** [12] to optimize the **Convolution Additive Token Mixer (CATM)** [13], we propose the **ECA-CATM** module, which achieves more controllable global contextual interaction while diminishing background interference on the feature representation of small face targets.
3. We devise a novel loss function termed **Focal-PIoUv2**, by incorporating the Focal mechanism [14] into the original Piou loss [15]. This loss expedites bounding box regression toward targets while suppressing the gradient interference imposed on the model by a high volume of low-quality predicted boxes.

2 Related Work

2.1 Dense Small-Scale Face Detection

We analyze the strengths and limitations of two mainstream object detectors, YOLO series and DETR series models, in the task of dense small face detection.

YOLO models leverage the local inductive bias of convolutional operators and high computational efficiency. They excel at extracting important local features for small targets and offer superior real-time performance. However, these models rely heavily

on NMS post-processing and possess relatively limited global context modeling. In dense detection scenarios, the NMS logic frequently suppresses overlapping targets as redundant anchor boxes, leading to a significant miss detection rate [16]. Although the recent YOLO26 achieves an NMS-free framework via a Consistent Dual Assignment strategy, it lacks collaborative inference among detection units, making it less effective at precisely delineating boundaries of overlapping targets.

DETR models achieve efficient global context modeling capability through self-attention and are inherently NMS-free. Therefore, they can accurately reason about spatial relationships using global context, accurately distinguishing overlapping targets. However, unconstrained global self-attention tends to introduce background clutter, causing the model to misidentify complex environmental textures as face targets. In addition, conventional DETR models suffer from high computational complexity, hindering real-time deployment. Nevertheless, RT-DETR concurrently attains high real-time performance and an NMS-free paradigm through its efficient hybrid encoder, manifesting substantial performance potential for the specific tasks addressed in our study.

2.2 Convolution Additive Token Mixer

In modern object detectors, the encoder's token mixer is central to global context modeling. Yet, conventional Multi-Head Self-Attention (MHSA) encounters several challenges in small face detection tasks. It struggles to extract fine-grained structural features of small targets. Furthermore, its pairwise dot-product across the entire feature map leads to the dilution of attention weights by complex environmental textures.

The CATM addresses these limitations by prioritizing noise reduction and local feature enhancement within a global context modeling framework. Central to CATM is the principle that effective global context modeling arises from synergistic information interactions across both spatial and channel domains. To implement this, the module employs an additive attention mechanism, with convolutions facilitating global interactions. This design yields more controllable global context information with suppressed noise. Crucially, CATM leverages the inherent local inductive bias of convolution to further strengthen the extraction of fine-grained structural details.

3 Method

3.1 Structure Overview

The overall architecture of the GFSM-DETR is shown in **Fig. 1**. Built upon the RT-DETR framework, the model adopts the lightweight MobilenetV4 [17] as its backbone to balance detection accuracy with inference speed. In GFSM-DETR, we incorporate the novel Focal-PloUv2 loss function and restructure the hybrid by integrating our proposed ECA-CATM and GFSM-Fusion.

The GFSM-Fusion replaces the original Fusion modules to amplify high-frequency textures of small targets and suppress background noise, thereby mitigating the loss of fine-grained features. Furthermore, to circumvent background noise interference inherent in standard self-attention, ECA-CATM supersedes the conventional MHSA-based

AIFI module, achieving more noise-resistant global context modeling. Finally, the Focal-PIoUv2 loss function replaces the GIoU loss in RT-DETR to more comprehensively suppress disruptive gradients from a high volume of low-quality predicted boxes and refine bounding box regression in complex environments.

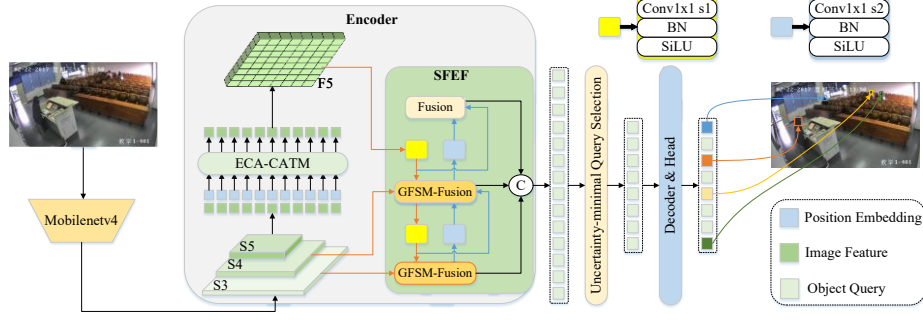


Fig. 1. The overall structure of our proposed GFSM-DETR

3.2 Gated Frequency-Spatial Modulation Fusion

Design Motivation and Theoretical Foundation of Frequency-Domain Analysis.

In complex environments, the subtle features required for detecting small face targets are often obscured by large-scale background clutter. Frequency-domain analysis offers an effective solution to this challenge. A series of studies [18,19,20,21] have validated that the efficacy of frequency-domain analysis in object detection stems from its ability to achieve accurate disentanglement and adaptive modulation of high-frequency target details and low-frequency background information. From the perspective of signal processing, the small-scale features of the target correspond to high-frequency signals, while the background is represented by low-frequency signals. Common spatial-domain optimizations often struggle to learn high-frequency features [18]. In contrast, performing frequency-domain modulation operations such as spectral convolutions enable adaptive weight assignment across different frequency bands, thereby enhancing the small-scale high-frequency features of targets while suppressing low-frequency background interference [19,20]. Furthermore, the downsampling used in the original Fusion module essentially functions as a low-pass filter [21]. The resulting information loss further motivates the use of active frequency enhancement strategies.

Inspired by these insights, we propose the GFSM-Fusion module, which integrates frequency-domain modulation with spatial-domain geometric modeling. By adaptively amplifying high-frequency target details and suppressing background noise, this module strengthens GFSM-DETR’s ability to detect overlapping and minute face targets.

Gated Frequency-Spatial Convolution. The GFSCov module serves as the technical core for achieving synergistic spatial-frequency enhancement. As shown in Fig. 2, it generates dynamic gating maps by modeling features in both domains simultaneously.

Initially, the input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ undergoes a 1×1 convolution to compress the channel dimensionality to $C/2$, providing a compact representation for the subsequent integration of spectral and spatial information.

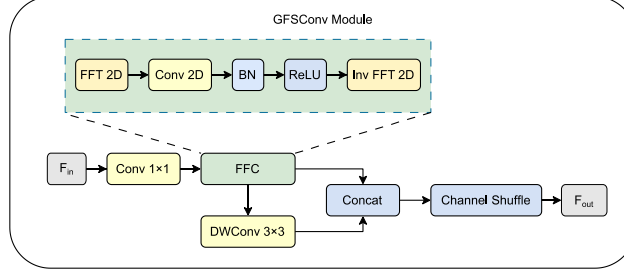


Fig. 2. Illustration of the GFSConv.

Frequency-Domain Analysis. The compressed feature map is then processed by a Fourier Unit, where a 2D Fast Fourier Transform (FFT) decomposes spatial features into physically meaningful spectral components, enabling the model to precisely characterize the spectral difference between target textures and background noise from a global perspective. Subsequently, a 1×1 convolution in the frequency domain serves as a learnable filter to perform dynamic spectral modulation, accurately recalibrating the gain of specific spectral components. This process amplifies the high-frequency small-scale essential features of face targets while suppressing background clutter. The modulated features are restored to the spatial domain via the Inverse Fast Fourier Transform (iFFT), yielding the frequency-modulated feature map $\mathbf{X}_f$.

$$\mathbf{X}_f = \text{iFFT}_{2D} \left(\text{Conv}_{\text{freq}} \left(\text{FFT}_{2D} \left(\text{Conv}_{1 \times 1}(\mathbf{X}) \right) \right) \right) \quad (1)$$

Spatial Refinement. To integrate global spectral priors into local geometric representations, $\mathbf{X}_f$ is processed through a 3×3 depthwise convolution (DWConv), generating the spatial refinement representation $\mathbf{X}_s$.

Joint Gating Map Generation. Subsequently, the frequency-modulated feature map $\mathbf{X}_f$ and the spatial refinement feature map $\mathbf{X}_s$ are concatenated channel-wise. A Channel Shuffle operation is applied to achieve interleaved fusion of the two different feature maps. The resulting joint gating map $\mathbf{G}_{fs}$ integrates global texture statistical characteristics in the frequency domain and the local feature information in the spatial domain.

$$\mathbf{G}_{fs} = \text{ChannelShuffle}(\text{Concat}(\mathbf{X}_f, \mathbf{X}_s)) \quad (2)$$

Feature Recalibration via Gating Mechanism. As illustrated in **Fig. 3**, within the GFSM-Fusion module, N Reparameterized blocks (RepBlocks) [23] form the core pathway for feature processing, which is responsible for extracting deep-level spatial feature representation $\mathbf{F}_{base}$. Using the joint gating map $\mathbf{G}_{fs}$, a gating recalibration is

performed on the primary feature path via an element-wise product. Incorporating a residual connection F_{res} , the final output is defined as follows:

$$F_{output} = (F_{base} \odot G_{fs}) + F_{res} \quad (3)$$

Utilizing collaborative frequency-spatial information, this method implements adaptive feature modulation to enhance fine-grained target features while attenuating background interference.

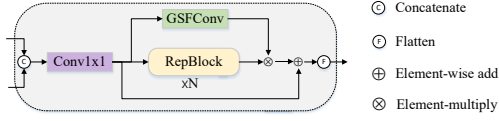


Fig. 3. Overall structure of the GFSM-Fusion module.

3.3 ECA-CATM

Motivation and Module Overview. While RT-DETR employs MHSA for global context modeling on the S_5 feature map, the global dot-product operation often leads to the dilution of critical small-target features by environmental textures in complex scenes.

To resolve these problems, we propose the ECA-CATM module as a superior alternative, as shown in Fig. 4. Based on an additive attention mechanism with linear time complexity, this module achieves controlled, noise-resistant global context modeling through sequential information interactions in spatial and channel domains. Crucially, the module leverages the local inductive bias of convolutional operators to enhance fine-grained features while preserving global modeling capability.

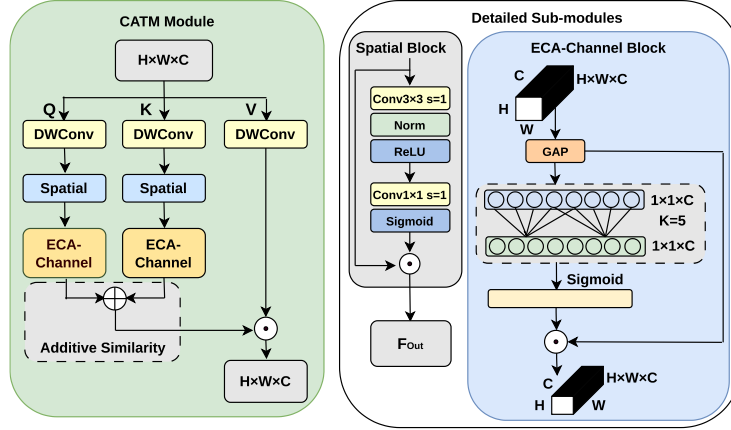


Fig. 4. Schematic of the ECA-CATM module.

Spatial Interaction Stage. As the initial step of feature processing, the spatial interaction stage aims to strengthen local target features and perform preliminary background

noise filtering. The input feature map $\mathbf{X}_{in} \in \mathbb{R}^{H \times W \times C}$ first passes through a 3×3 DWConv layer to extract local structural features. This effectively utilizes the convolutional inductive bias to compensate for the inherent limitations of global attention in capturing micro-scale details. Subsequently, a 1×1 convolution layer is adopted to compress the channel dimension to 1, followed by a Sigmoid activation to generate a 2D spatial attention map. Finally, this map is applied to the input feature map $\mathbf{X}_{in}$ to suppress background clutter and emphasize the critical features of small face targets, obtaining the spatially enhanced representation $\mathbf{X}_{spatial}$.

$$\mathbf{X}_{spatial} = \text{Sigmoid} \left(\text{Conv}_{1 \times 1} \left(\text{ReLU} \left(\text{BN}(\text{DWConv}_{3 \times 3}(\mathbf{X}_{in})) \right) \right) \right) \odot \mathbf{X}_{in} \quad (4)$$

ECA-Based Channel Interaction Stage. Channel information interaction is carried out on the spatially enhanced feature map $\mathbf{X}_{spatial}$. CATM’s channel interaction method employs dimensionality reduction, which inevitably destroys the delicate detail representations essential for identifying small face targets. To address this, we integrate the ECA mechanism, which avoids dimensionality reduction to preserve important small-scale features. Specifically, ECA first aggregates global spatial contextual information via Global Average Pooling (GAP). Then a 1D convolution layer with a kernel size of k is utilized to implement cross-channel information interaction, where the weight of each channel is determined by its k neighboring channels. This strategy avoids the loss of critical feature information and retains the vital fine-grained feature representations. The calibrated channel-enhanced feature map $\mathbf{X}_{channel}$ is formulated as follows:

$$\mathbf{X}_{channel} = \text{Sigmoid} \left(\text{Conv1D}_k \left(\text{GAP}(\mathbf{X}_{spatial}) \right) \right) \odot \mathbf{X}_{spatial} \quad (5)$$

Additive Similarity and Feature Aggregation. To model global correlations without the computational intensity and noise sensitivity of MHSA, this stage employs an additive similarity function.

$$\text{Sim}(\mathbf{Q}, \mathbf{K}) = \psi(\mathbf{Q}) + \psi(\mathbf{K}) \quad (6)$$

Here, the ψ denotes the global context modeling process involving information interaction across the spatial and channel domains. In addition, as displayed in **Fig. 4**, the Query ($\mathbf{Q}$), Key ($\mathbf{K}$) and Value ($\mathbf{V}$) are derived by processing the input feature map $\mathbf{X}$ of the ECA-CATM module through three independent DWConv layers respectively.

The resulting similarity weights are applied to the value branch $\mathbf{V}$ via an element-wise product. To further bolster the feature representation, a 1×1 DWConv layer is introduced to refine the integrated contextual semantics, followed by a final linear projection to complete the feature mapping.

$$\mathbf{F}_{out} = \text{Linear}(\text{DWConv}_{1 \times 1}(\text{Sim}(\mathbf{Q}, \mathbf{K}) \odot \mathbf{V})) \quad (7)$$

3.4 Focal-PIoUv2 Bounding Box Regression Loss Function

Design Motivation and GIoU Limitations. In the scenario of dense small face detection, the number of target faces is large, yet high-quality predicted boxes with tight fitting and high overlap with target boxes are far fewer than low-quality ones. The large cumulative gradient generated by massive low-quality predicted boxes severely misleads the optimization direction of the model training. Additionally, the GIoU loss employed in the original RT-DETR tends to excessively expand predicted boxes, and it degenerates to IoU when predictions and target boxes have an inclusion relationship, losing the ability to refine their relative positions [15]. These defects collectively result in the model's slow convergence and overall performance limitations.

To address these challenges, we propose the Focal-PIoUv2 loss function, which synergistically suppresses gradient interference from diverse low-quality samples while providing precise geometric constraints for direct regression.

Distance-Guided Direct Regression. To circumvent the problems of blind expansion of predicted boxes and degradation in the process of GIoU, we introduce the PIoU loss. Specifically, it constructs a penalty factor P , which incorporates the boundary distance between the predicted and target box as well as the target scale. With this penalty term, the loss function can effectively guide the predicted boxes to match the target boxes more directly and accurately. The formulation is as follows:

$$L_{PIoU} = 2 - IoU - e^{-P^2} \quad (8)$$

Synergistic Suppression of Low-Quality Predicted boxes. During training, low-quality predicted boxes are mainly categorized into two types: those with a quite low IoU with the target, and those that have a certain IoU with the target but exhibit extremely abnormal shapes. Relying solely on one existing suppression mechanism is insufficient to simultaneously filter out a large number of these two types of low-quality predicted boxes. Therefore, in the proposed Focal-PIoUv2 loss function, we establish the synergistic interaction between the Focal mechanism and the non-monotonic attention function of PIoUv2. This synergy achieves a more comprehensive suppression of disruptive gradients generated by low-quality predicted boxes.

Building upon the PIoU, the PIoUv2 loss constructs a non-monotonic attention function based on the penalty factor P to suppress low-quality boxes. This approach excels at restricting predicted boxes with excessive spatial offsets and abnormal shapes. However, it exhibits limited suppression capability for predicted boxes with low IoU. In the following formula, the expression for q is $q = e^{-P}$.

$$L_{PIoUv2} = 3 \cdot (\lambda q) \cdot e^{-(\lambda q)^2} \cdot L_{PIoU} \quad (9)$$

The Focal mechanism introduces an IoU-based weighting factor W_{Focal} to specifically attenuate the gradient contributions of low-quality predicted boxes with negligible IoU with the targets. However, it is inherently limited in constraining those predictions characterized by anomalous geometric shapes.

$$W_{\text{Focal}} = IoU^\alpha \quad (10)$$

By constructing the Focal-PIoUv2 loss function, we achieve a synergistic complementarity between the two mechanisms. Under this dual-constraint framework, the model simultaneously accounts for the overlap degree, spatial positional relationships, and geometric shape consistencies between the predicted and target boxes. This collaborative approach effectively suppresses disruptive gradients originating from various types of low-quality predicted boxes, ensuring that the model always focuses on the regression of high-quality predicted boxes. Consequently, it improves the regression stability of the model and its overall detection performance in practical scenarios.

$$L_{\text{Focal-PIoUv2}} = L_{\text{PIoUv2}} \cdot W_{\text{Focal}} = 3 \cdot (\lambda q) \cdot e^{-(\lambda q)^2} \cdot IoU^\alpha \cdot L_{\text{PIoU}} \quad (11)$$

4 Experiments

4.1 Datasets

To evaluate the performance of our proposed method in complex environments with high-density face targets, we conduct experiments on two publicly available benchmark datasets: SCUT-Head [24] and Brainwash [25].

SCUT-Head is released by South China University of Technology and contains 4,405 images with 111,241 annotated small face targets. Part A of the dataset is derived from classroom surveillance videos, featuring a high density of face targets and significant scale differences among them. Part B is sourced from complex scene internet images, encompassing various perspectives, backgrounds and face poses.

Brainwash is constructed by Stanford University and includes 11,789 images with 91,146 small face targets annotated. All the images are from video frames captured by a fixed indoor surveillance camera in a café at different times, ensuring diverse lighting conditions despite the consistent perspective. The targets are characterized by minute scales, high density, and significant pose variations. These characteristics lead to severe mutual occlusion and heavy background interference.

4.2 Implementation Details and Evaluation Metrics

The proposed model is implemented using PyTorch 2.4.1 framework and accelerated by CUDA 12.1 on an Ubuntu 20.04 operating system. The hardware environment consists of an NVIDIA GeForce RTX 4080 SUPER GPU (32GB memory) and 17 vCPU Intel(R) Xeon(R) Platinum 8470Q CPU. For data preprocessing, all input images are standardized to a resolution of 640×640 pixels. The training process is conducted for 150 epochs with a batch size of 16. To ensure experimental reproducibility, the random seed is fixed at 0, and deterministic training is enabled with 8 data-loading worker threads. We choose the AdamW optimizer and a cosine annealing learning rate scheduler for optimization, with an initial learning rate of 0.01, a momentum value of 0.937, and a weight decay of 5×10^{-4} . For the other configurations of the models, the IoU

threshold is set to 0.7, and the maximum number of model detection results is limited to 300. In addition, in the Focal-PIoUv2 loss, λ is set to 1.3 and the α is set to 0.5.

We utilize the following metrics to evaluate the accuracy of our model in identifying small face targets: Precision (P), Recall (R), F1-Score, and mAP_{50} . Average Precision (AP) is defined as the area under the Precision-Recall curve, while mean Average Precision (mAP) represents the mean of AP values across all categories. Specifically, mAP_{50} signifies that a detection is considered correct if the IoU between the predicted and target boxes is greater than or equal to 0.5.

$$AP = \int_0^1 P(R) dR \quad (12)$$

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (13)$$

Model efficiency is quantified using Params (Parameters) and GFLOPs (Giga Floating-point Operations). Params denotes the total volume of learnable parameters in the model, indicating the model’s memory occupancy. GFLOPs represents the number of floating-point operations required for a single inference of the model, measuring its computational complexity.

4.3 Comparative Experiment

To comprehensively evaluate the effectiveness of our proposed model, we select several mainstream object detectors as benchmarks, including the YOLO series (v5, v8, v11 and 26s/m), classic detectors (Faster R-CNN [26] and SSD [27]), and the original RT-DETR model. All models are evaluated under the identical experimental environment and hyperparameter configuration described in Section 4.2.

The primary challenges of the SCUT-Head dataset lie in its extremely high target density and drastic scale variation. As indicated in **Table 1**, based on a holistic evaluation of the performance metrics, the top three performing baseline models on this dataset are YOLOv8m, YOLOv11m and YOLO26m. Experimental results demonstrate that our proposed model exhibits significant advantages. Compared to these leading baselines, our model achieves a gain of over 3.9% in Recall and an improvement of over 2.5% in both F1-score and mAP_{50} , while maintaining comparable Precision. More importantly, compared with those of the top-performing YOLO models, our model reduces Params and GFLOPs by over 24.5% and 18.7%. This suggests that by optimizing the feature extraction architecture, our method more effectively addresses the challenges posed by high target density and complex backgrounds in SCUT-Head. It achieves high detection accuracy while exhibiting great lightweight characteristics and real-time performance.

The Brainwash dataset primarily focuses on multi-pose face targets, complex illumination interference, and severe mutual occlusion among abundant small-scale face targets. The top-performing baseline models on this dataset are YOLOv5m, YOLOv11m and YOLO26m. As presented in **Table 2**, our proposed model maintains its leading position. With similar Precision, our model improves Recall by over 2.0%,

F1-score by 0.6% and mAP_{50} by 2.0%. The reductions in Params and GFLOPs are consistent with the margins observed on the previous dataset. Despite the severe target overlap and environmental interference, the significant improvement in Recall and mAP_{50} directly validates the effectiveness of our model in reducing the missed detection issue of overlapping faces.

Analyzing the results across both datasets, our model consistently achieves the highest detection accuracy across various complex small face detection scenarios. The substantial improvements in Recall and mAP_{50} confirm that our approach effectively resolves the missed detection of overlapping targets. Furthermore, the remarkable reduction in memory footprint and computational cost enables it to serve as a high-accuracy, low-cost solution for deployment and real-time execution on edge and mobile devices.

Table 1. The comparative experimental results on the SCUT-Head dataset.

Models	P(%)	R(%)	F1(%)	mAP50 (%)	Params(M)	GFLOPs(G)
SSD	82.0	67.0	73.7	84.5	25.2	31.0
Faster R-CNN	86.5	79.0	82.6	86.0	28.3	474.1
YOLOv5s	94.1	85.1	89.8	91.1	9.1	23.8
YOLOv5m	94.2	87.3	91.1	92.2	25.0	64.2
YOLOv8s	94.2	85.9	89.8	91.5	11.1	28.4
YOLOv8m	94.8	88.0	91.3	92.4	25.8	78.7
YOLOv11s	94.1	85.1	89.7	91.0	9.4	21.3
YOLOv11m	94.3	87.9	91.5	92.6	20.0	67.6
YOLO26s	94.2	84.8	89.7	91.0	9.4	20.5
YOLO26m	94.5	87.6	91.4	92.4	20.4	67.8
RT-DETR	92.0	88.6	90.5	92.7	15.7	54.5
Ours	94.6	92.5	93.6	95.2	15.1	54.9

Table 2. The comparative experimental results on the Brainwash dataset.

Models	P(%)	R(%)	F1(%)	mAP50 (%)	Params(M)	GFLOPs(G)
SSD	84.4	72.5	78.0	83.1	25.2	31.0
Faster R-CNN	88.5	80.4	84.3	87.0	28.2	474.1
YOLOv5s	93.9	84.9	89.2	91.9	9.1	23.8
YOLOv5m	95.3	88.3	91.7	93.7	25.0	64.2
YOLOv8s	93.8	85.2	89.3	92.1	11.1	28.4
YOLOv8m	95.8	87.3	91.4	93.4	25.8	78.7
YOLOv11s	93.7	86.1	89.7	92.5	9.4	21.3
YOLOv11m	95.7	88.9	92.2	93.6	20.0	67.6
YOLO26s	93.9	86.1	89.9	92.6	9.4	20.5
YOLO26m	95.2	88.7	91.8	93.5	20.4	67.8
RT-DETR	93.2	88.7	90.9	93.3	15.7	54.5
Ours	94.9	90.9	92.8	95.7	15.1	54.9

4.4 Ablation Study

To systematically investigate the contributions of the proposed optimization modules to the model’s performance, we conduct comprehensive ablation experiments on the SCUT-Head dataset. The results of the ablation study are shown in **Table 3**.

Replacing the original AIFI module with ECA-CATM yields significant improvements in both the computational efficiency and detection accuracy of the model. The sole integration of ECA-CATM improves Precision from 92.0% to 93.1%, while simultaneously reducing the Params from 15.7M to 15.3M. This performance gain confirms that ECA-CATM not only suppresses background noise but also enhances the accurate capture of local features for small face targets, thereby preventing the model from misidentifying environmental textures as face targets.

The introduction of Focal-PIoUv2 gains improvements across all primary metrics: Precision, Recall, F1-score and mAP_{50} increased by 1.7%, 1.8%, 1.5% and 1.1%, respectively. This result validates that the novel loss accurately guides bounding box regression. Furthermore, it alleviates the distribution imbalance between high-quality and low-quality predicted boxes, compelling the model to prioritize the regression of high-quality boxes, which ultimately elevates the overall detection accuracy.

The contribution of the GFSM-Fusion strategy to model performance is particularly substantial. When introduced as an independent component, it achieves gains of 1.4%, 2.2%, 1.6% and 1.2% in Precision, Recall, F1-score and mAP_{50} , respectively. More importantly, GFSM-Fusion delivers consistent performance gains across all module combinations, providing a stable contribution ranging from 1.2% to 2.2%, without exhibiting typical diminishing marginal utility effect in multi-module integration. This phenomenon underscores the superior capability of the frequency-spatial collaborative enhancement strategy in extracting and amplifying high-frequency textures and critical features, thereby effectively addressing missed detections caused by severe target overlapping or small target size in complex environments.

In summary, the introduction of any proposed component or their combination into the RT-DETR results in consistent performance increments. Furthermore, compared to the RT-DETR, our GFSM-DETR realizes marked gains of 2.6% in Precision, 3.9% in Recall, 3.1% in F1-score and 2.5% in mAP_{50} . The results of the ablation experiment not only verify the effectiveness of each optimization scheme but also highlight the profound synergistic effect between the modules.

Table 3. The ablation study results on the SCUT-Head dataset.

Models	CATM	GFSM	Loss	P(%)	R(%)	F1(%)	mAP50 (%)	Params(M)	GFLOPs(G)
1.RT-DETR				92.0	88.6	90.5	92.7	15.7	54.5
2	✓			93.1	89.6	91.3	93.3	15.3	54.4
3			✓	93.7	90.4	92.0	93.8	15.7	54.5
4		✓		93.4	90.8	92.1	93.9	15.5	55.0
5	✓		✓	93.9	90.9	92.3	94.1	15.3	54.4
6		✓	✓	93.7	91.6	92.6	94.4	15.5	55.0
7	✓	✓		93.7	91.7	92.7	94.4	15.1	54.9
8	✓	✓	✓	94.6	92.5	93.6	95.2	15.1	54.9

4.5 Visual Analysis

To assess the practical performance of GFSM-DETR, we conduct a visualization analysis using representative samples from the SCUT-Head and Brainwash datasets. As illustrated in **Fig. 5**, these chosen cases are quite challenging for the models.

The first row of **Fig. 5** presents an image from the SCUT-Head dataset. In this image, targets in the rear classroom exhibit tiny scales, high density and heavy mutual occlusion, with substantial clutter caused by background desks and chairs. These issues are particularly evident in the white-box region. Upon close observation, we can see that RT-DETR misses a small target overlapping with other faces and incorrectly classifies a background region as a face. By comparison, our GFSM-DETR accurately detects all these targets. These observations confirm that GFSM-DETR can accurately detect overlapping small targets and effectively distinguish them from the background.

The cases in the second and third rows are from the Brainwash dataset. Both cases contain a human whose body is visible, but whose face is more than half occluded by walls or lights. RT-DETR misses both targets, whereas our GFSM-DETR detects them correctly. These results indicate that our GFSM-DETR is highly sensitive to fine-scale features and possesses strong capabilities in modeling contextual information.



Fig. 5. Visualization of detection results. Green bounding boxes indicate ground truth targets, while red bounding boxes represent the model’s detections. The white bounding boxes indicate the areas we need to magnify for detailed observation.

5 Conclusion

To address the challenges of target density, mutual occlusion, background interference and drastic scale variations in complex environments, this paper proposes a lightweight

and efficient GFSM-DETR based on the RT-DETR framework. The designed GFSM-Fusion strategy employs a frequency-spatial collaborative modulation mechanism to accurately capture and amplify the high-frequency critical features of small face targets, significantly bolstering the model's perceptual capability for overlapping and micro-scale faces. Furthermore, by introducing the ECA-CATM module, we leverage an additive attention mechanism to achieve more controllable global context modeling, which effectively suppresses interference from complex backgrounds. Finally, the Focal-PIoUv2 loss function is utilized to optimize the bounding box regression process, enabling the model to focus more on the regression of high-quality prediction boxes and thereby further enhancing the model's localization accuracy.

Experimental results on two datasets validate that GFSM-DETR outperforms existing mainstream object detectors, providing a high-accuracy and low-cost solution for real-time dense face detection on edge and mobile devices. Despite the remarkable progress achieved, there remains potential for further performance optimization. Our future research will focus on enhancing the detection performance of lightweight models in extreme dynamic environments and their cross-scenario transferability, continuously improving the reliability and practical application value of the model in complex scenarios.

References

1. Feng, Y., Yu, S., Peng, H., et al.: Detect faces efficiently: a survey and evaluations. *IEEE Trans. Biometrics, Behavior, Identity Sci.* 4(1), 1–18 (2022)
2. Deepal, D.A.A., De Silva, R., Ariyaratne, M.K.A., et al.: Faces unveiled: a deep dive into modern face detection and recognition techniques. *Int. J. Res. Comput.* 4(i), 48–81 (2025)
3. Jocher, G., et al.: Ultralytics YOLOv5. Zenodo (2020). <https://doi.org/10.5281/zenodo.7347926>, last accessed 2026/03/26
4. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. <https://github.com/ultralytics/ultralytics> (2023), last accessed 2026/03/26
5. Ultralytics: YOLO11: real-time object detection and beyond. <https://docs.ultralytics.com/models/yolo11> (2024), last accessed 2026/03/26
6. Sapkota, R., Cheppally, R.H., et al.: YOLO26: key architectural enhancements and performance benchmarking for real-time object detection. *arXiv preprint arXiv:2509.25164* (2025)
7. Carion, N., Massa, F., Synnaeve, G., et al.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) *ECCV 2020. LNCS*, vol. 12347, pp. 213–229. Springer, Cham (2020)
8. Zhu, X., Su, W., Lu, L., et al.: Deformable DETR: deformable transformers for end-to-end object detection. In: *International Conference on Learning Representations (ICLR) 2021* (2021)
9. Zhang, H., Li, F., Xu, H., et al.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: *International Conference on Learning Representations (ICLR) 2023* (2023)
10. Zong, Z., Song, G., Liu, Y.: DETRs with collaborative training for object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2023*, pp. 6751–6761 (2023)

11. Lv, W., Zhao, Y., Xu, S., et al.: DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16965–16974 (2024)
12. Wang, Q., Wu, B., Zhu, P., et al.: ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11534–11542 (2020)
13. Zhang, T., Li, L., Zhou, Y., et al.: CAS-ViT: convolutional additive self-attention vision transformers for efficient mobile applications. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
14. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)
15. Liu, C., Wang, K., Li, Q., et al.: Powerful-IoU: more straightforward and faster bounding box regression loss with a non-monotonic focusing mechanism. *Neural Networks* 169, 11–41 (2024)
16. Bodla, N., Singh, B., Chellappa, R., et al.: Soft-NMS: improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5561–5569 (2017)
17. Qin, D., Leichner, C., Delagrangue, M., et al.: MobileNetV4: universal models for the mobile ecosystem. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
18. Sun, H., Wang, R., Li, Y., et al.: SET: spectral enhancement for tiny object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025, pp. 4713–4723 (2025)
19. Rao, Y., Zhao, W., Zhu, Z., et al.: Global filter networks for image classification. *Adv. Neural Inf. Process. Syst.* 34, 980–993 (2021)
20. Xu, K., Qin, M., Sun, F., et al.: Learning in the frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020, pp. 1740–1749 (2020)
21. Jiang, L., Dai, B., Wu, W., et al.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2021, pp. 13919–13929 (2021)
22. Sun, K., Xiao, B., Liu, D., et al.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703 (2019)
23. Ding, X., Zhang, X., Ma, N., et al.: RepVGG: making VGG-style ConvNets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13733–13742 (2021)
24. Peng, D., Li, Z., Fu, J., et al.: SCUT-HEAD: a new benchmark for head detection. In: Proceedings of the 24th International Conference on Pattern Recognition (ICPR), pp. 1328–1333. IEEE (2018)
25. Stewart, R., Andriluka, M., Ng, A.Y.: End-to-end people detection in crowded scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 2325–2333 (2016)
26. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(6), 1137–1149 (2017)
27. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016)