Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 8, 2026: 3111-3129
ISSN:2302-8149
e-ISSN:2540-9719

Adaptive Mix-Gated Multi-Attention Fusion in YOLOv11s for Dense Person Detection: A Systematic Evaluation of Dual and Triple Attention Configurations

[1]Aya Ahmed*, [2]Karam Abdullah
[1,2]Computer Science Department, University of Mosul, Al-Thaqafa Street, Mosul 41002, Iraq
*e-mail: aya.24esp1@student.uomosul.edu.iq

(***received***: 14 July 2026, ***revised***: 24 July 2026, ***accepted***: 27 July 2026)

Abstract

Person detection in a crowded environment is very challenging due to excessive occlusion, scale variation, and heavy overlap between individuals. This paper proposes a lightweight Mix-Gated attention framework to enhance the feature representation capacity of YOLOv11s for dense person detection. The proposed framework incorporates two complementary channel-wise and spatial attention mechanisms under a flexible channel-wise gating strategy, while preserving the original YOLOv11s architecture and real-time inference ability. Unlike previous studies, which primarily employed either a single attention mechanism or fixed combinations of multiple attention modules, the proposed framework introduces an adaptive Mix-Gated feature selection mechanism that dynamically learns the contribution of complementary attention modules through trainable channel-wise gating. Furthermore, this study systematically evaluates ten dual and triple attention configurations under identical training and evaluation conditions, providing a comprehensive and fair comparison of their effectiveness. this study evaluated the performance on four popular benchmark datasets - CrowdHuman, WiderPerson, MOT17Det and MOT20Det - using Precision, Recall, F1-score, mAP@50 and mAP@50-95 as evaluation metrics. The experimental results show that all proposed Mix-Gated configurations outperform the baseline YOLOv11s model. Mix-Gated-CBAM+SE+SA has the best detection accuracy and efficient inference performance among Mix-Gated configurations. These results indicate that adaptive fusion of complementary attention mechanisms can improve the feature representation and person detection performance in crowded scenes while not sacrificing computational efficiency.

Keywords: *dense person detection, YOLOv11s, mix-gated attention, CBAM, SE, ECA, spatial attention, crowd analysis.*

1. Introduction

With the continuous growth of the global population and the rapid increase in large public gatherings, people detection is one of the most important research topics in computer vision. People detection is the basis for many intelligent applications such as video surveillance [1], [2], public safety monitoring [3], autonomous driving [4], crowd management [5], multi-object tracking [1], human behavior analysis [6], and vision-based navigation and smart robotic systems [7]. The effectiveness of these applications depends on detection models capable of delivering high accuracy and real-time performance in complex visual situations, and accurate detection algorithms are a major goal of today's computer vision research.

 ISSN:2302-8149
e-ISSN:2540-9719

Even with recent advancements, accurate person detection in crowded environments is still a challenge due to the high degree of occlusions, scale variations, complex backgrounds and dense object distributions that are often missed and localized in the wrong direction [8], [9], [10], [11].

To overcome these limitations, object detection has been greatly advanced with the use of deep learning with the traditional handcrafted feature extraction techniques as Haar Cascade, Histogram of Oriented Gradients (HOG)[12], and Scale-Invariant Feature Transform (SIFT) [13] with CNNs that can automatically learn hierarchical feature representations from data. Deep learning models have shown better detection accuracy, robustness and generalization to complex visual environments [14], [15].

Among modern detectors, the YOLO family has become popular due to the good detection accuracy and computational efficiency. While YOLOv11 improves feature extraction and inference speed, the lightweight YOLOv11s model still suffers from poor recall and localization accuracy in highly crowded scenes. This is because the single attention pathway cannot efficiently take advantage of the complementary channel-wise and spatial information [16], [17]. More recently, attention mechanisms [18] have become one of the most effective ways to enhance feature representation in deep neural networks by making it possible to emphasize relevant features while suppressing irrelevant information. Representative attention modules, including the Convolutional Block Attention Module (CBAM)[19], Squeeze-and-Excitation (SE)[20], Efficient Channel Attention (ECA)[21], and Spatial Attention (SA)[22], have shown remarkable improvements in the detection of small, densely packed, and partially occluded objects. It has also been recently shown that incorporating attention mechanisms in YOLO-based detectors significantly improves pedestrian detection performance in crowded environments[23], [24], [25], [26], [27], [28].

However, most of the studies have focused on one attention mechanism or a few attention combinations and do not consider the complementary advantages of combined attention mechanisms. They propose a lightweight Mix-Gated YOLOv11s framework to adaptively fuse multiple channel and spatial attention mechanisms with a simple gating strategy. the evaluate ten attention configurations, including six dual-attention and four triple-attention, on CrowdHuman, MOT17Det, MOT20Det, and WiderPerson benchmark datasets with Precision, Recall, F1-score, mAP@50, and mAP@50–95. find that all the Mix-Gated YOLOv11s versions perform better than the baseline YOLOv11s while Mix-Gated-CBAM+SE+SA has the best overall performance. This shows that adaptive multi-attention fusion can be implemented to detect dense people without affecting real-time inference efficiency.

The main contributions of this study include the proposal of a lightweight Mix-Gated framework based on YOLOv11s for pedestrian detection in densely populated environments. The proposed framework integrates multiple channel-wise and spatial attention mechanisms through an adaptive Mix-Gated fusion strategy. Furthermore, this study presents a systematic comparison of ten dual- and triple-attention configurations under identical training and evaluation settings. The proposed models are comprehensively evaluated on four challenging benchmark datasets, namely CrowdHuman, MOT17Det, MOT20Det, and WiderPerson, using Precision, Recall, F1-score, mAP@50, and mAP@50–95 as evaluation metrics. The experimental results demonstrate that the integration of complementary attention mechanisms significantly enhances feature representation and pedestrian detection performance while preserving the computational efficiency of the YOLOv11s architecture.

 ISSN:2302-8149
e-ISSN:2540-9719

2. Literature Review

Person detection in crowded environments remains a major problem in computer vision because of occlusion, scale variation, dense overlap, and complex backgrounds that decrease detection accuracy and increase missed detections [29]. Deep learning, particularly the YOLO family, has become the leading framework for real-time object detection because it balances between accuracy and computational efficiency. he latest YOLOv11 architecture further improves feature extraction while still being efficient, thereby making it good for surveillance and crowd analysis applications [30]. As a result, recent works have been focused on improving YOLO-based detectors by improving the backbone, feature fusion, and attention mechanisms for crowded-scene detection [17]. Table 1 summarizes the attention-based approaches from different perspectives with their adopted attention mechanisms, key findings, and remaining drawbacks.

Tabel 1 Summary of representative studies on attention-enhanced deep learning models.

Authors / Model	Attention Mechanism	Advantages	Limitations
YOLOv5s [31]	ECA + SE	The BaseSE-ECA model demonstrated outstanding accuracy, reaching a precision of 96.69% and a mAP of 98.4%.	Focused on autonomous driving rather than crowded person detection; no adaptive multi-attention fusion.
CNN [32]	CBAM – SE	The CBAM-enhanced CNN achieved 98.6 % accuracy, improving upon the baseline CNN's 92.08 %, with a sensitivity of 98.3 % and specificity of 97.9 %. The SE-integrated CNN followed with 96.25 % accuracy, demonstrating superior feature recalibration. ResNet50 with CBAM attained 93.32 % accuracy.	Medical image classification; not applicable to person detection.
YOLOv7 [33]	SE-CBAM	the proposed SE-CBAM-YOLOv7 improves detection accuracy by 0.5% and the mAP value by 1.7% compared to YOLOv7	Evaluated aircraft targets rather than pedestrian detection in crowded scenes.
U-Net [34]	SA + Dilated ECA + CBAM + SE	Achieved 84.30% Dice coefficient, outperforming the baseline U-Net and improving segmentation of small and irregular pulmonary nodules.	not designed for person detection or adaptive Mix-Gated fusion.
YOLOv5n, YOLOv5m[35]	CBAM-ECA	YOLOv5n + CBAM achieved the best performance with Precision = 84.4%, Recall = 71.1%, F1-score = 77.2%, and mAP@0.5 = 77.8%, outperforming the baseline YOLOv5n.	Applied only to landslide detection; evaluated CBAM and ECA individually without adaptive multi-attention fusion; not tested for crowded person detection.
YOLOv8 [36]	SimAM + SE + CBAM + ECA	Achieved 79.8% mAP@0.5, improving 4.7% over YOLOv8n. Reached an F1-score of 75%, while reducing parameters by 34.4% and computational complexity by 41.4%, making it suitable for edge-device deployment.	Evaluated only on the NEU-DET steel surface defect dataset. Focuses on industrial defect detection rather than person detection or crowded scenes and does not employ adaptive Mix-Gated fusion for combining multiple attention mechanisms.
YOLOv8[37]	HCA (Hybrid Context Attention):	Outperformed seven state-of-the-art models across Precision, Recall, F1-score, mAP@0.5, and mAP@0.5:0.95. Improved mAP@0.5 by	not combinations adaptive Mix-Gated fusion.

 ISSN:2302-8149
e-ISSN:2540-9719

	integrates Channel Attention + Spatial Attention + Multi-scale Context Attention + Global Semantic Attention	12.89% on the Crowd dataset and 8.29% on the KITTI dataset compared with YOLOv8s, while maintaining the same number of parameters. Demonstrated superior detection in dense occlusion, small objects, and complex backgrounds.

feature representation and object detection performance. However, most approaches rely on a single attention mechanism or a small number of attention combinations with fixed feature fusion strategies. The most widely adopted mechanisms are SE, ECA, SA, and CBAM, which provide complementary channel-wise and spatial feature enhancement [19], [20], [21], [22]. To address this issue, the systematically test various dual- and triple-attention configurations in the lightweight YOLOv11s architecture with an adaptive Mix-Gated fusion strategy for dense person detection.

3. Research Method

This section describes the research methodology adopted in this study:

Research Framework

The proposed research framework is to test the effectiveness of integrating multiple attention mechanisms in the lightweight YOLOv11s architecture for dense person detection. Figure 1 shows the four benchmark datasets (CrowdHuman, WiderPerson, MOT17Det, MOT20Det) which are preprocessed before training. The proposed approach improves the baseline YOLOv11s model by adding Mix-Gated attention modules to the backbone layers of C3k2 while preserving the original neck and detection head. Ten dual and triple-attention configurations are tested under identical training conditions. Attention outputs are fused with a Softmax-based learnable weighting strategy and then combined with the backbone features through the proposed channel-wise Mix-Gated fusion mechanism. Then all the models are evaluated according to Precision, Recall, F1-score, mAP@50, mAP@50–95, and inference efficiency to find the best attention configuration.

 ISSN:2302-8149
e-ISSN:2540-9719

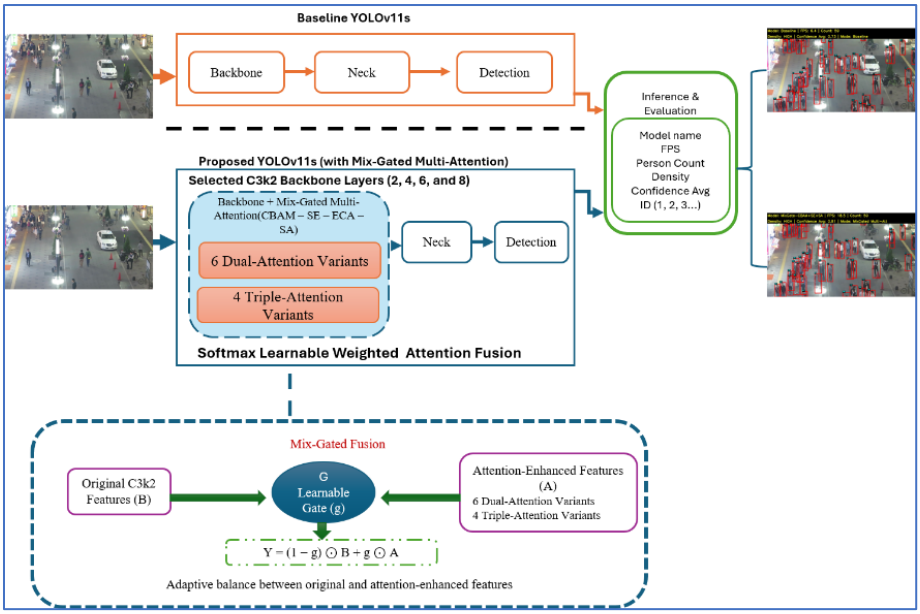

Figure 1 Overall research workflow of the proposed Mix-Gated multi-attention framework based on YOLOv11s for dense person detection.

Algorithm 1 Proposed Mix-Gated YOLOv11s Framework for Dense Person Detection

Require :Benchmark datasets D = {CrowdHuman, MOT17Det, MOT20Det, WiderPerson}

Pretrained YOLOv11s model

Attention configurations C = {C1, C2, ..., C10}

Ensure: Performance comparison between the baseline YOLOv11s model and the proposed Mix-Gated attention models

1: Train and evaluate the baseline YOLOv11s model
2: *Save baseline performance metrics*
3: *for each attention configuration* $Ci \in C$ *do*
4: *Initialize a new YOLOv*11*s model*
5: Select C3k2 backbone layers {2, 4, 6, 8}
6: *for each selected C*3*k*2 *layer do*
7: *Extract baseline feature map Fbase*
8: *Apply the attention modules defined in Ci*
9: Generate attention − enhanced feature maps
10: *FuseattentionoutputsusingSoftmaxlearnableweights*
11: Generate the channel − wise gating vector G
12: Compute the Mix − Gated output
13: Replace the original C3k2 output with Fout
14: *end for*
15 : *Train and validate the modified YOLOv*11*s model*
16: Evaluate the model using Precision, Recall, F1 − score, mAP@50, mAP@50
17: *Save the evaluation results*14: *end for*
18: *end for*
19: *Compare all proposed models with the baseline YOLOv*11*s*
20: *Return the best − performing attention configuration*

 ISSN:2302-8149
e-ISSN:2540-9719

Baseline YOLOv11s Model

The baseline model used in this work is YOLOv11s. This is the reference model to compare and contrast to the proposed Mix-Gated attention framework. YOLOv11s is a good choice because it provides a good balance between the accuracy of the detection, computational efficiency, and inference speed and is suitable for real-time dense person detection. The model consists of 3 main parts: the backbone for feature extraction, the neck for multi-scale feature fusion, and the detection head for bounding boxes, confidence scores, and class labels. No attention mechanisms are incorporated in the baseline model, which is the basis of the comparison to the proposed framework.

Attention Configurations

To evaluate the effectiveness of many attention mechanisms in YOLOv11s, this study investigate ten attention configurations under the same training and evaluation conditions. Unlike previous works that focus on a single attention mechanism or a fixed combination of attention mechanisms, the proposed framework compares multiple complementary dual-attention and triple-attention configurations that are in the same lightweight YOLOv11s architecture. The configurations that are considered include six dual-attention and four triple-attention models and are summarized in Table 2.

Table 2 Evaluated dual- and triple-attention configurations used in the proposed framework

NO	Attention Configuration	Type
1	CBAM + ECA	Dual-Attention
2	CBAM + SE	Dual-Attention
3	CBAM + SA	Dual-Attention
4	ECA + SE	Dual-Attention
5	ECA + SA	Dual-Attention
6	SE + SA	Dual-Attention
7	CBAM + ECA + SE	Triple-Attention
8	CBAM + ECA + SA	Triple-Attention
9	CBAM + SE + SA	Triple-Attention
10	ECA + SE + SA	Triple-Attention

The selected attention combinations are chosen to leverage the complementary features of SE, ECA, SA, and CBAM which enhance channel-wise, spatial or combined feature representations. All configurations were incorporated into the same YOLOv11s architecture using our Mix-Gated framework and trained with the same datasets, preprocessing procedures, hyperparameter settings, optimization strategy and evaluation measures. So, any performance differences can be directly attributed to the adopted attention configuration.

Attention Modules

To enhance feature representation within the YOLOv11s backbone, this study adopts four complementary attention mechanisms: Squeeze-and-Excitation (SE), Efficient Channel Attention (ECA), Spatial Attention (SA), and the Convolutional Block Attention Module (CBAM). these mechanisms improve feature learning by emphasizing informative channel and spatial information while suppressing less relevant features. Although they share the common objective of enhancing feature representation, each mechanism follows a different attention strategy.

Specifically, SE recalibrates channel responses by modeling inter-channel dependencies through global average pooling, whereas ECA efficiently captures local channel interactions using lightweight

 ISSN:2302-8149
e-ISSN:2540-9719

one-dimensional convolution without dimensionality reduction. In contrast, SA focuses on identifying informative spatial regions to improve feature localization, while CBAM sequentially combines channel and spatial attention to produce more discriminative feature representations. Owing to their complementary characteristics, these attention mechanisms are integrated within the proposed Mix-Gated framework, which systematically evaluates multiple dual-attention and triple-attention configurations to determine the most effective feature enhancement strategy for dense person detection.

Softmax Learnable Weighted Fusion

To combine the outputs of multiple attention mechanisms this study adopts a Softmax-normalized learnable weighted fusion strategy. Instead of a conventional fusion method which concatenates or equally averages the outputs of multiple attention mechanisms, the proposed strategy allows the network to learn the relative importance of each attention branch during training. So, the contribution of each attention mechanism is adapted to the feature representation from the data. For each attention configuration, the selected attention modules are applied in parallel to the same feature map extracted from the selected C3k2 block. Let B denote the original feature map extracted from the baseline YOLOv11s backbone. Each attention branch generates an enhanced feature map, denoted by A_k, where k represents the attention branch within the selected configuration. Depending on the evaluated model, the number of attention branches is either two (dual-attention) or three (triple-attention). As shown in Equation (1), the attention-enhanced feature map is obtained by combining the outputs of the selected attention branches using learnable Softmax weights:

$$A = \sum_{k=1}^{N} \alpha_k A_k \qquad (1)$$

where **A** denotes the fused attention feature map, $\mathbf{A_k}$ is the output of the k-th attention module, $\boldsymbol{\alpha_k}$ is the normalized learnable weight assigned to that branch, and **N** represents the number of attention modules in the selected configuration.

The attention weights are computed using the Softmax function, as defined in Equation (2):

$$\alpha_k = \frac{e^{\lambda_k}}{\sum_{j=1}^{N} e^{\lambda_j}} \qquad (2)$$

where $\boldsymbol{\lambda_k}$ is the learnable parameter associated with the k-th attention branch. Equation (3) ensures that the normalized attention weights satisfy.

$$\sum_{k=1}^{N} \alpha_k = 1 \qquad (3)$$

the relative importance of each attention mechanism during training, such that the more effective mechanisms are assigned larger weights, while the contribution of less important ones is reduced. As a result, the fused feature map A can exploit the complementary characteristics of the different attention mechanisms before being passed to the Mix-Gated module.

Figure 2 illustrates the learnable weighted fusion mechanism based on Softmax, where the selected attention branches are processed in parallel and their outputs are then combined using adaptive learnable weights to produce an enhanced feature map that is subsequently used in the Mix-Gated Fusion stage.

 ISSN:2302-8149
e-ISSN:2540-9719

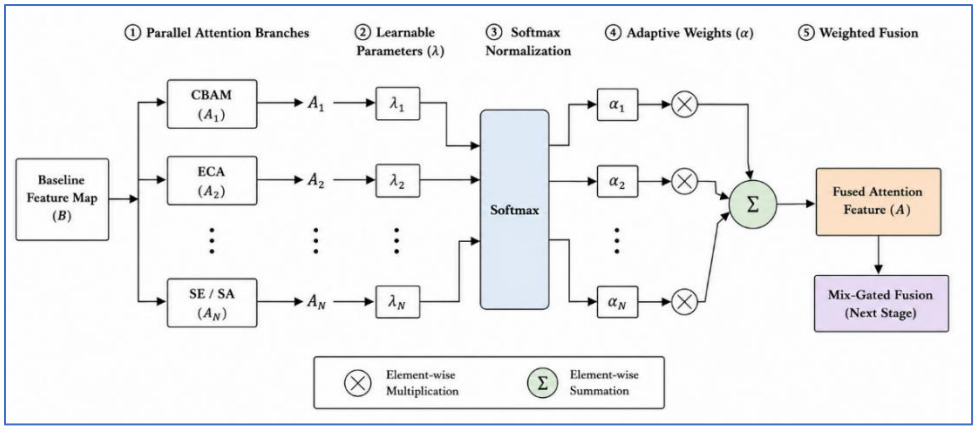

Figure 2 Softmax-based learnable weighted fusion mechanism. the selected attention modules are processed in parallel, and their outputs are adaptively combined using learnable softmax weights to generate the fused attention feature map (A) before being forwarded to the Mix-Gated Fusion module.

Channel-wise Mix-Gated Fusion

After obtaining the attention-enhanced feature map (A) from the Softmax learnable weighted fusion stage, the proposed framework applies a Channel-wise Mix-Gated Fusion module to adaptively combine the original backbone features with the attention-enhanced features. Unlike conventional fusion methods that employ fixed fusion rules, the proposed approach introduces a learnable channel-wise gate that dynamically determines the contribution of each feature source during training.

Let B denote the original feature map extracted from the selected C3k2 block, and let A represent the attention-enhanced feature map generated by the Softmax learnable weighted fusion stage. As defined in Equation (4), a channel-wise gate vector g is generated using the Sigmoid activation function to produce gate values ranging from 0 to 1, as expressed by

$$g = \text{Sigmoid}\,(G) \qquad (4)$$

where G denotes the learnable gating parameters, and g represents the normalized channel-wise gate vector after the Sigmoid activation.

Using the gate values obtained from Equation (4), the final fused feature map is computed according to Equation (5):

$$F = (1 - g) \odot B + g \odot A \qquad (5)$$

where: F denotes the final fused feature map. B represents the original feature map extracted from the selected C3k2 block of the YOLOv11s backbone. A denotes the attention-enhanced feature map obtained from the Softmax learnable weighted fusion stage. g is the normalized channel-wise gate vector that adaptively controls the contribution of the original and attention-enhanced feature maps. $\odot$ denotes the element-wise multiplication operator.

The proposed Channel-wise Mix-Gated Fusion mechanism enables the network to adaptively balance the original and attention-enhanced feature representations according to the extracted visual information. When the gate values are close to 0, the model relies primarily on the original backbone features. Conversely, when the gate values approach 1, greater emphasis is placed on the attention-enhanced features. For intermediate gate values, both feature representations contribute proportionally, allowing the network to preserve important spatial information while simultaneously exploiting the discriminative features learned by the attention mechanisms. This adaptive fusion

 ISSN:2302-8149
e-ISSN:2540-9719

strategy improves feature representation without introducing modifications to the neck or detection head of the YOLOv11s architecture.

The overall architecture of the proposed Channel-wise Mix-Gated Fusion module is illustrated in Figure 3, where the original feature map (B) and the attention-enhanced feature map (A) are adaptively combined through the learnable channel-wise gate (g) to generate the final fused feature map (F) before being forwarded to the subsequent layers of the YOLOv11s backbone.

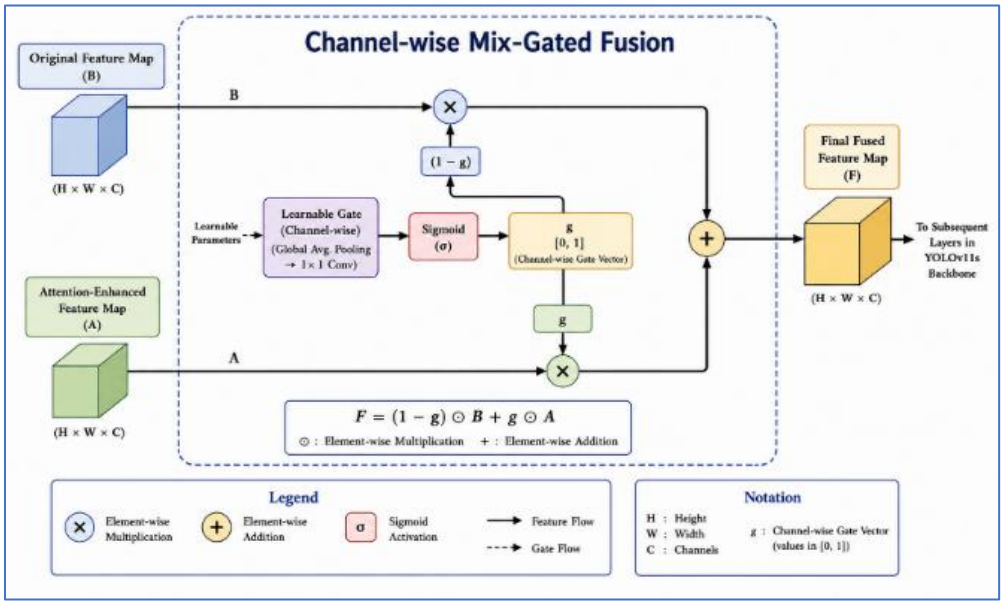

Figure 3 illustrates the overall architecture of the proposed Channel-wise Mix-Gated Fusion module, in which the original feature map (B) and the attention-enhanced feature map (A) are adaptively fused using a learnable channel-wise gate vector (g) to generate the final fused feature map (F). the resulting feature representation is subsequently forwarded to the subsequent layers of the YOLOv11s backbone for further feature extraction and object detection.

Injection Strategy

the proposed Mix-Gated attention module was selectively integrated into the backbone of the YOLOv11s architecture by replacing the feature representation of some C3k2 blocks and preserving the overall network architecture. Instead of inserting module into every layer of the backbone, chose to inject it into four selected C3k2 layers: layers 2, 4, 6, and 8. This approach was adopted to improve feature representation at a range of semantic levels without adding any additional computational burden.

The selected C3k2 layers capture complementary feature information throughout the backbone. The earlier layers preserve fine-grained spatial details that are crucial for detecting small or partially occluded pedestrians, while the deeper layers encode richer semantic information that improves the discrimination of overlapping individuals in densely populated scenes. By enhancing feature representations at different depths, our framework can improve both localization accuracy and feature discrimination without modifying the remaining components of the detector.

On each selected C3k2 block, the original feature map is first extracted and fed into the attention configuration. Depending on the evaluation model, we will apply two or three attention mechanisms at the same time to generate multiple attention-enhanced feature maps. These features are combined

 ISSN:2302-8149
e-ISSN:2540-9719

using the proposed Softmax learnable weighted fusion strategy and fused with the original backbone features using the proposed Channel-wise Mix-Gated Fusion module. The final fused feature map then replaces the original output of the corresponding C3k2 block before being propagated to the subsequent backbone layers.

To ensure a fair comparison with the baseline YOLOv11s model, the neck, feature pyramid, and detection head are the same in all experiments. Therefore, any improvements in performance can be attributed to the proposed feature enhancement strategy instead of architectural changes in other parts of the detector. The overall injection strategy in the proposed framework is shown in Figure 4 The Mix-Gated attention module is added to the selected C3k2 layers of the YOLOv11s backbone while keeping the original detection pipeline.

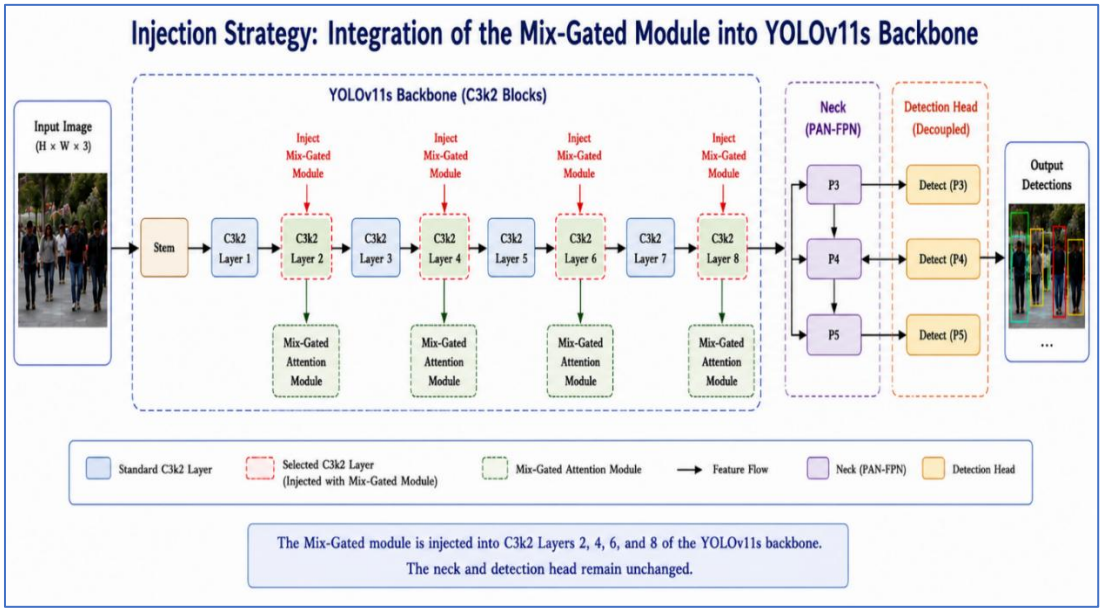

Figure 4 Injection strategy of the proposed Mix-Gated attention module into the selected C3k2 layers of the YOLOv11s backbone.

Inference and Evaluation Pipeline

After training, all models were compared using the same inference pipeline to ensure a fair comparison between the baseline YOLOv11s model and the proposed Mix-Gated attention variants. Precision, Recall, F1-score, mAP@50, and mAP@50-95 were evaluated. The framework also generates automatic person counting and separates each image into low, medium, or high crowd density levels for descriptive purposes. All experiments were performed using the same experimental setting, so no performance differences were seen only with the proposed Mix-Gated attention strategy. Finally, the ten attention configurations were compared to the baseline YOLOv11s model to determine which one is the best for dense person detection.

4. Results and Analysis

This section presents the experimental evaluation of the proposed Mix-Gated multi-attention-enhanced YOLOv11s framework through extensive experiments on challenging benchmark datasets for dense person detection.

Dataset

To fully evaluate the proposed Mix-Gated attention-enhanced YOLOv11s framework, performed experiments on four popular benchmark datasets for dense person detection: CrowdHuman,

 ISSN:2302-8149
e-ISSN:2540-9719

WiderPerson, MOT17Det, and MOT20Det. These datasets were selected to account for different real-world scenarios with varying crowd sizes, severe occlusions, scale variations, complex backgrounds, and dynamically changing scene conditions.

CrowdHuman[38] is specifically designed for highly crowded environments with an overwhelming number of people overlapping and a high level of occlusion, so it is suitable for evaluating detection robustness in a challenging environment. WiderPerson[39] is based on pedestrians of different sizes, particularly small and distant individuals, and is a good benchmark to assess scale-invariant detection performance. MOT17Det[40] has dynamic surveillance scenes with variations in motion, viewpoint, and illumination, and hence can be used in real tracking situations. MOT20Det[41] is very dense urban crowds and is one of the most challenging benchmarks for person detection due to its very high crowd density and severe occlusion. Taken together, these benchmark datasets provide a perfect environment to evaluate our framework under various crowd characteristics and visual conditions and demonstrate the robustness and generalization of the proposed framework.

Experimental Setup

All experiments were conducted on a Windows workstation equipped with an NVIDIA GeForce RTX 4050 Laptop GPU (6 GB GDDR6). The proposed framework was implemented using Python 3.10.18 in the Spyder integrated development environment, based on PyTorch 2.5.1 with CUDA 12.1 and the Ultralytics YOLO framework (v8.3.221). The lightweight YOLOv11s model was selected as the baseline because it provides an effective balance between detection accuracy, computational efficiency, and real-time inference.

The experiments were performed on four benchmark datasets: CrowdHuman, WiderPerson, MOT17Det, and MOT20Det. All datasets were converted into the YOLO annotation format, and the detection task was restricted to the person class (Class ID = 0). Each dataset was divided into training, validation, and testing subsets according to its official configuration. The same dataset format and evaluation protocol were adopted for all experiments to ensure a fair comparison.

To evaluate the effectiveness of the proposed framework, the baseline YOLOv11s model was compared with ten Mix-Gated attention configurations, including six dual-attention models (CBAM+ECA, CBAM+SA, CBAM+SE, ECA+SA, ECA+SE, and SE+SA) and four triple-attention models (CBAM+ECA+SE, CBAM+ECA+SA, CBAM+SE+SA, and ECA+SE+SA). In all configurations, the selected attention modules were inserted into the same C3k2 layers of the YOLOv11s backbone. Their outputs were first fused using Softmax-normalized learnable weights and then adaptively combined with the original backbone features through the proposed channel-wise Mix-Gated fusion mechanism. The neck and detection head remained unchanged for all models.

To ensure a fair comparison, all models were trained using identical training settings. Training was performed for a maximum of 460 epochs with early stopping enabled (patience = 15). All input images were resized to 640 × 640 pixels using a batch size of 8. The same data augmentation strategy, including HSV augmentation, translation, scaling, mosaic, mixup, horizontal flipping, and copy-paste augmentation, was applied to every model. Unless otherwise specified, all remaining hyperparameters followed the default Ultralytics YOLO training configuration.

During evaluation, all models were tested using the same inference settings. Detection performance was measured using Precision, Recall, F1-score, mAP@50, mAP@50–95. The confidence threshold was set to 0.15, the IoU threshold to 0.55, and the maximum number of detections to 500 for all experiments.

 ISSN:2302-8149
e-ISSN:2540-9719

Evaluation of the Proposed Framework

The proposed Mix-Gated attention framework was compared to the baseline YOLOv11s model by comparing the ten Mix-Gated attention configurations on CrowdHuman, WiderPerson, MOT17Det, and MOT20Det benchmark datasets. Performance was measured using Precision, Recall, F1-score, mAP@50, mAP@50–95. The quantitative and qualitative analyses were carried out to identify the most effective attention configuration and evaluate its robustness, localization accuracy, and inference efficiency in dense crowd scenes.

Quantitative Performance Comparison

In this subsection, the compared the baseline YOLOv11s model to the ten Mix-Gated attention configurations. We analyze the experimental results obtained in the four benchmark datasets to identify the best attention configuration and assess the trade-off between detection accuracy and computational efficiency.

Performance on the MOT20Det Dataset

Table 3 presents the quantitative comparison between the baseline YOLOv11s model and the ten proposed Mix-Gated attention configurations on the MOT20Det dataset, one of the most challenging benchmarks due to its extremely dense crowds and severe occlusions.

Table 3 Quantitative results on the MOT20Det dataset.

Model	Precision	Recall	mAP50	mAP50–95	F1-score
Baseline	94.44	87.74	93.84	64.17	0.9097
Mix-Gated-CBAM+ECA	96.91	92.31	96.93	73.79	0.9456
Mix-Gated-CBAM+SA	97.04	92.48	97.00	74.78	0.9470
Mix-Gated-CBAM+SE	97.32	92.23	96.93	74.75	0.9470
Mix-Gated-ECA+SA	97.77	92.97	97.24	77.78	0.9531
Mix-Gated-ECA+SE	94.66	85.41	92.32	54.83	0.8980
Mix-Gated-SE+SA	97.32	91.95	96.81	74.29	0.9456
Mix-Gated-CBAM+ECA+SE	96.98	92.45	96.95	74.23	0.9466
Mix-Gated-CBAM+ECA+SA	97.39	93.13	97.32	76.89	0.9521
Mix-Gated-CBAM+SE+SA	97.92	94.00	97.72	80.70	0.9592
Mix-Gated-ECA+SE+SA	97.18	92.52	97.06	75.17	0.9479

model in most evaluation metrics, confirming the effectiveness of adaptive attention fusion for dense person detection. Among all evaluated models, Mix-Gated-CBAM+SE+SA achieved the best overall performance, obtaining the highest Precision (97.92%), Recall (94.00%), mAP@50 (97.72%), and mAP@50–95 (80.70%), with an F1-score of 0.9592.

Performance on the CrowdHuman Dataset

Table 4 presents the quantitative comparison between the baseline YOLOv11s model and the ten proposed Mix-Gated attention configurations on the CrowdHuman dataset, one of the most challenging benchmarks due to its extremely dense crowds and severe occlusions.

Table 4 Quantitative results on the CrowdHuman dataset

Model	Precision	Recall	mAP50	mAP50–95	F1-score
Baseline	66.93	46.47	52.60	27.12	0.5485

 ISSN:2302-8149
e-ISSN:2540-9719

Mix-Gated-CBAM+ECA	81.28	57.80	67.42	38.78	0.6756
Mix-Gated-CBAM+SA	81.06	58.94	68.09	39.53	0.6825
Mix-Gated-CBAM+SE	80.20	57.53	66.49	37.78	0.6700
Mix-Gated-ECA+SA	79.62	56.15	65.08	36.57	0.6586
Mix-Gated-ECA+SE	78.98	55.50	64.40	36.14	0.6519
Mix-Gated-SE+SA	79.47	56.61	65.45	36.98	0.6612
Mix-Gated-CBAM+ECA+SE	81.15	59.27	68.41	39.79	0.6851
Mix-Gated-CBAM+ECA+SA	81.01	59.76	69.11	40.41	0.6907
Mix-Gated-CBAM+SE+SA	81.53	60.41	69.47	40.75	0.6940
Mix-Gated-ECA+SE+SA	81.41	59.11	68.38	39.60	0.6849

Overall, all proposed Mix-Gated attention configurations outperformed the baseline YOLOv11s model. Among them, Mix-Gated-CBAM+SE+SA achieved the best overall performance, with Precision of 81.53%, Recall of 60.41%, mAP@50 of 69.47%, mAP@50–95 of 40.75%, and an F1-score of 0.6940, compared with 66.93%, 46.47%, 52.60%, 27.12%, and 0.5485, respectively, for the baseline model. These results demonstrate the effectiveness of the proposed adaptive Mix-Gated attention framework for dense person detection.

Performance on the MOT17Det Dataset

Table 5 presents the quantitative comparison between the baseline YOLOv11s model and the ten proposed Mix-Gated attention configurations on the MOT17Det dataset, one of the most challenging benchmarks due to its extremely dense crowds and severe occlusions.

Table 5 Quantitative results on the MOT17Det dataset.

Model	Precision	Recall	mAP50	mAP50–95	F1-score
Baseline	90.90	78.63	88.31	62.44	0.8432
Mix-Gated-CBAM+ECA	95.18	86.70	94.63	73.51	0.9074
Mix-Gated-CBAM+SA	95.48	88.27	95.68	74.79	0.9173
Mix-Gated-CBAM+SE	95.15	87.31	94.85	74.56	0.9106
Mix-Gated-ECA+SA	94.35	86.86	94.62	74.84	0.9045
Mix-Gated-ECA+SE	94.75	86.38	94.59	73.97	0.9037
Mix-Gated-SE+SA	95.21	86.57	94.71	73.54	0.9069
Mix-Gated-CBAM+ECA+SE	95.53	88.06	95.48	74.51	0.9164
Mix-Gated-CBAM+ECA+SA	96.47	90.18	96.54	78.30	0.9322
Mix-Gated-CBAM+SE+SA	97.18	91.05	96.68	79.36	0.9402
Mix-Gated-ECA+SE+SA	95.48	87.88	95.12	74.54	0.9152

Overall, all proposed Mix-Gated attention configurations outperformed the baseline YOLOv11s model. Among them, Mix-Gated-CBAM+SE+SA achieved the best overall performance, with Precision of 97.18%, Recall of 91.05%, mAP@50 of 96.68%, mAP@50–95 of 79.36%, and an F1-score of 0.9402, compared with 90.90%, 78.63%, 88.31%, 62.44%, and 0.8432, respectively, for the baseline model. These results demonstrate the effectiveness of the proposed Mix-Gated framework in dynamic surveillance scenes.

 ISSN:2302-8149
e-ISSN:2540-9719

Performance on the WiderPerson Dataset

Table 6 presents the quantitative comparison between the baseline YOLOv11s model and the ten proposed Mix-Gated attention configurations on the WiderPerson dataset, one of the most challenging benchmarks due to its extremely dense crowds and severe occlusions.

Table 6 Quantitative results on the WiderPerson dataset.

Model	Precision	Recall	mAP50	mAP50–95	F1-score
Baseline	77.68	64.81	75.20	44.10	0.7066
Mix-Gated-CBAM+ECA	83.78	78.72	86.89	58.27	0.8117
Mix-Gated-CBAM+SA	83.81	79.23	87.24	58.82	0.8146
Mix-Gated-CBAM+SE	84.16	77.58	86.17	57.77	0.8074
Mix-Gated-ECA+SA	83.65	76.27	85.33	56.67	0.7979
Mix-Gated-ECA+SE	83.13	76.11	84.92	56.39	0.7947
Mix-Gated-SE+SA	83.43	77.11	85.86	57.18	0.8015
Mix-Gated-CBAM+ECA+SE	83.97	79.34	87.54	59.05	0.8158
Mix-Gated-CBAM+ECA+SA	84.11	80.00	87.94	59.71	0.8216
Mix-Gated-CBAM+SE+SA	84.20	80.05	87.92	59.85	0.8211
Mix-Gated-ECA+SE+SA	83.87	79.05	87.21	58.99	0.8139

Overall, all Mix-Gated variants outperformed the baseline YOLOv11s model. Mix-Gated-CBAM+SE+SA achieved the best results, improving Precision, Recall, mAP@50, mAP@50–95, and F1-score from 77.68%, 64.81%, 75.20%, 44.10%, and 0.7066 to 84.20%, 80.05%, 87.92%, 59.85%, and 0.8211, confirming the effectiveness of the proposed framework.

Qualitative Performance Comparison

Figures 5–8 present qualitative comparisons between the baseline YOLOv11s model and the best-performing Mix-Gated configuration, demonstrating improved person detection in crowded scenes. As shown in Figure 5, the proposed model improves detection completeness on the CrowdHuman dataset by reducing missed detections under severe occlusion.

 ISSN:2302-8149
e-ISSN:2540-9719

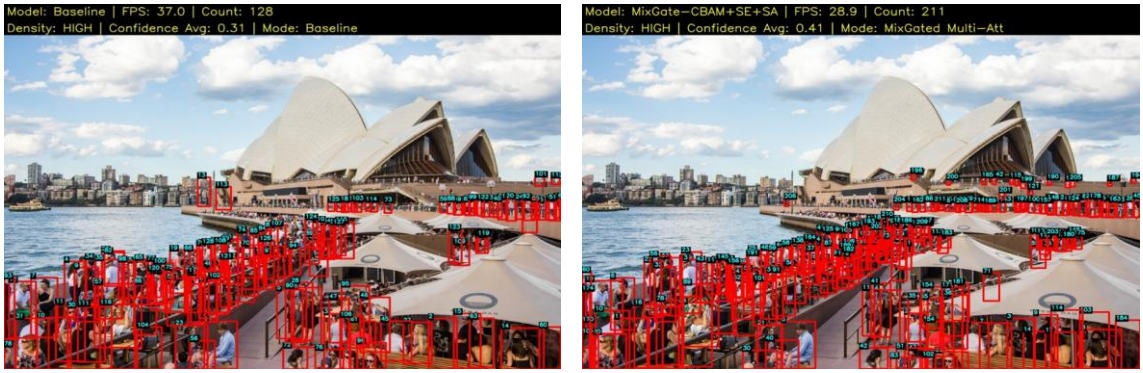

Figure 5 Qualitative comparison of detection results on the CrowdHuman dataset: (a) Baseline YOLOv11s; (b) Proposed Mix-Gated-CBAM+SE+SA.

Figure 6 demonstrates that the proposed model detects small and overlapping pedestrians more reliably on the WiderPerson dataset, producing more stable bounding boxes than the baseline model.

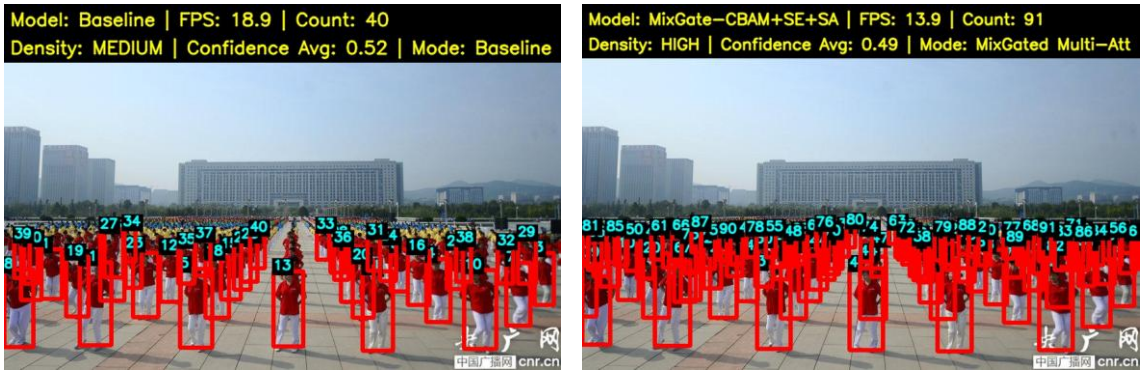

Figure 6 Qualitative comparison of detection results on the WiderPerson dataset: (a) Baseline YOLOv11s; (b) Proposed Mix-Gated-CBAM+SE+SA.

In Figure 7, the proposed model achieves more stable detections in the dynamic MOT17Det scenes affected by motion blur and partial occlusion, reducing missed and merged detections.

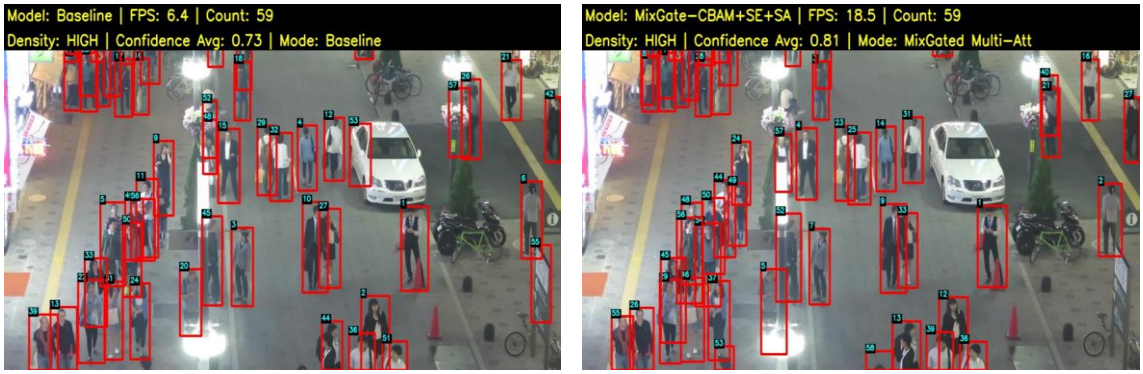

Figure 7 Qualitative comparison of detection results on the MOT17Det dataset: (a) Baseline YOLOv11s; (b) Proposed Mix-Gated-CBAM+SE+SA.

 ISSN:2302-8149
e-ISSN:2540-9719

Finally, Figure 8 illustrates the superiority of the proposed model on the highly crowded MOT20Det dataset by improving localization accuracy, reducing merged detections, and enhancing the detection of heavily occluded pedestrians.

Figure 8 Qualitative comparison of detection results on the MOT20Det dataset: (a) Baseline YOLOv11s; (b) Proposed Mix-Gated-CBAM+SE+SA.

Overall, the qualitative observations are consistent with the quantitative results, confirming that the proposed Mix-Gated framework improves person detection in crowded environments while preserving the real-time inference capability of YOLOv11s.

5. Conclusion

This study proposed a lightweight Mix-Gated attention framework to improve dense person detection based on YOLOv11s. Ten dual- and triple-attention configurations were systematically evaluated on the CrowdHuman, WiderPerson, MOT17Det, and MOT20Det datasets. The results showed that all proposed configurations outperformed the baseline model, with Mix-Gated-CBAM+SE+SA achieving the best overall performance in terms of Precision, Recall, F1-score, mAP@50, and mAP@50–95 while preserving computational efficiency compared with other attention-enhanced variants. These findings demonstrate that adaptive fusion of complementary attention mechanisms effectively enhances feature representation and improves person detection in crowded environments. Future work will investigate the proposed framework on larger real-world datasets and extend it to multi-object tracking applications.

References

[1] S. Abba, A. M. Bizi, J.-A. Lee, S. Bakouri, and M. L. Crespo, '*Real-Time Object Detection, Tracking, and Monitoring Framework for Security Surveillance Systems*', *Heliyon*, Vol. 10, No. 15, p. e34922, Aug. 2024, DOI: 10.1016/j.heliyon.2024.e34922.

[2] P. Raja, D. B K, R. T, V. Ravi, and N. S. Alghamdi, '*A Context-Aware Multi-Modal Generative Adversarial Network for Real-Time Anomaly Detection in Video Surveillance*', *Peer-to-Peer Netw. Appl.*, Vol. 19, No. 1, p. 29, Jan. 2026, DOI: 10.1007/s12083-025-02134-1.

[3] Y. Chen, J. Li, E. Blasch, and Q. Qu, *'Future Outdoor Safety Monitoring: Integrating Human Activity Recognition with the Internet of Physical–Virtual Things*', *Applied Sciences*, Vol. 15, No. 7, p. 3434, Mar. 2025, DOI: 10.3390/app15073434.

[4] M. Contreras, A. Jain, N. P. Bhatt, A. Banerjee, and E. Hashemi, '*A Survey on 3D Object Detection in Real Time for Autonomous Driving*', *Front. Robot. AI*, Vol. 11, p. 1212070, Mar. 2024, DOI: 10.3389/frobt.2024.1212070.

 ISSN:2302-8149
e-ISSN:2540-9719

[5] A. M. Alasmari, N. S. Farooqi, and Y. A. Alotaibi, '*Recent Trends in Crowd Management using Deep Learning Techniques: A Systematic Literature Review*', *J. Umm Al-Qura Univ. Eng.Archit.*, Vol. 15, No. 4, pp. 355–383, Dec. 2024, DOI: 10.1007/s43995-024-00071-3.

[6] S. Essahraui *et al.*, '*Human Behavior Analysis: A Comprehensive Survey on Techniques, Applications, Challenges, and Future Directions*', *IEEE Access*, Vol. 13, pp. 128379–128419, 2025, DOI: 10.1109/ACCESS.2025.3589938.

[7] E. A. Rodríguez-Martínez, W. Flores-Fuentes, F. Achakir, O. Sergiyenko, and F. N. Murrieta-Rico, '*Vision-based Navigation and Perception for Autonomous Robots: Sensors, SLAM, Control Strategies, and Cross-Domain Applications—A Review*', *Eng*, Vol. 6, No. 7, p. 153, Jul. 2025, DOI: 10.3390/eng6070153.

[8] C. Chen, J. Li, Z. Shuai, Y. Wang, and Y. Wang, '*A Lightweight Optimization Framework for Real-Time Pedestrian Detection in Dense and Occluded Scenes*', *Mech. SCI.*, Vol. 16, No. 2, pp. 877–886, Nov. 2025, DOI: 10.5194/ms-16-877-2025.

[9] S. Xing, F. Wang, and H. Wang, '*A Pedestrian Detection Model based on YOLO for Dense Scenes*', *Optoelectron. Lett.*, Vol. 22, No. 4, pp. 229–235, Apr. 2026, DOI: 10.1007/s11801-026-4274-2.

[10] Z. Hu, W. Niu, and S. Mo, '*CRD-YOLO: A High-Accuracy Real-Time Crowded Pedestrian Detection Algorithm*', *SIViP*, Vol. 20, No. 3, p. 159, Mar. 2026, DOI: 10.1007/s11760-026-05220-w.

[11] W. Sheng, M. Liu, X. Li, and M. Zhang, '*OLODN: An Efficient Lightweight People Detection Method for Occlusion and Crowding Scenarios*', *IET Image Processing*, Vol. 19, No. 1, p. e13325, Jan. 2025, DOI: 10.1049/ipr2.13325.

[12] N. Dalal and B. Triggs, '*Histograms of Oriented Gradients for Human Detection*', in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA: IEEE, 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.

[13] D. G. Lowe, '*Distinctive Image Features from Scale-Invariant Keypoints*', *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, Nov. 2004, DOI: 10.1023/B:VISI.0000029664.99615.94.

[14] E. Edozie, A. N. Shuaibu, U. K. John, and B. O. Sadiq, '*Comprehensive Review of Recent Developments in Visual Object Detection based on Deep Learning*', *Artif Intell Rev*, Vol. 58, No. 9, p. 277, Jun. 2025, DOI: 10.1007/s10462-025-11284-w.

[15] V. Pagire, M. Chavali, and A. Kale, '*A Comprehensive Review of Object Detection with Traditional and Deep Learning Methods*', *Signal Processing*, Vol. 237, p. 110075, Dec. 2025, DOI: 10.1016/j.sigpro.2025.110075.

[16] A. A. Murat and M. S. Kiran, '*A Comprehensive Review on YOLO Versions for Object Detection*', *Engineering Science and Technology, an International Journal*, Vol. 70, p. 102161, Oct. 2025, DOI: 10.1016/j.jestch.2025.102161.

[17] R. Sapkota *et al.*, '*YOLO Advances to its Genesis: A Decadal and Comprehensive Review of the You Only Look Once (YOLO) series*', *Artif Intell Rev*, Vol. 58, No. 9, p. 274, Jun. 2025, DOI: 10.1007/s10462-025-11253-3.

[18] M.-H. Guo *et al.*, '*Attention Mechanisms in Computer Vision: A Survey*', *Comp. Visual. Med.*, Vol. 8, No. 3, pp. 331–368, Sep. 2022, DOI: 10.1007/s41095-022-0271-y.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, '*CBAM: Convolutional Block Attention Module*', in *Computer Vision – ECCV 2018*, Vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y.

 ISSN:2302-8149
e-ISSN:2540-9719

Weiss, Eds, in Lecture Notes in Computer Science, Vol. 11211. , Cham: Springer International Publishing, 2018, pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1.

[20] J. Hu, L. Shen, and G. Sun, '*Squeeze-and-Excitation Networks*', 2018.

[21] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, '*ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks*', in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 11531–11539. DOI: 10.1109/CVPR42600.2020.01155.

[22] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, '*An Empirical Study of Spatial Attention Mechanisms in Deep Networks*', in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 6687–6696. DOI: 10.1109/ICCV.2019.00679.

[23] M.-H. Bae, S.-W. Park, J. Park, S.-H. Jung, and C.-B. Sim, '*YOLO-RACE: Reassembly and Convolutional Block Attention for Enhanced Dense Object Detection*', *Pattern Anal Applic*, Vol. 28, No. 2, p. 90, Jun. 2025, DOI: 10.1007/s10044-025-01471-4.

[24] X. Chen, J. Li, H. Zhang, J. Gong, H. Wu, and O. Postolache, '*Vehicle Navigation-Oriented Object Distance Estimation via a Squeeze and Excitation Channel Attention Mechanism-Enhanced YOLO Module*', 2026, *SSRN*. DOI: 10.2139/ssrn.6143995.

[25] Z. Chen, S. Tian, L. Yu, L. Zhang, and X. Zhang, '*An Object Detection Network based on YOLOv4 and Improved Spatial Attention Mechanism*', *IFS*, Vol. 42, No. 3, pp. 2359–2368, Feb. 2022, DOI: 10.3233/JIFS-211648.

[26] C.-T. Chien, R.-Y. Ju, K.-Y. Chou, E. Xieerke, and J.-S. Chiang, '*YOLOv8-AM: YOLOv8 based on Effective Attention Mechanisms for Pediatric Wrist Fracture Detection*', *IEEE Access*, Vol. 13, pp. 52461–52477, 2025, DOI: 10.1109/ACCESS.2025.3549839.

[27] Z. Tang, J. Lu, Z. Chen, F. Qi, and L. Zhang, '*Improved Pest-YOLO: Real-Time Pest Detection based on Efficient Channel Attention Mechanism and Transformer Encoder*', *Ecological Informatics*, Vol. 78, p. 102340, Dec. 2023, DOI: 10.1016/j.ecoinf.2023.102340.

[28] L. Wu, X. Li, P. Ma, and Y. Cai, '*Research on a Dense Pedestrian-Detection Algorithm based on an Improved YOLO11*', *Future Internet*, Vol. 17, No. 10, p. 438, Sep. 2025, DOI: 10.3390/fi17100438.

[29] D. Chavan and A. Purohit, '*Machine Learning Techniques for Crowd Counting: A Survey*', *IJCA*, Vol. 185, No. 37, pp. 1–8, Oct. 2023, DOI: 10.5120/ijca2023923176.

[30] I. Beqqali Hassani, S. Benhida, N. Lamii, K. Oqaidi, A. Ouiddad, and S. Ghiadi, *'From YOLO V1 to YOLO V11: Comparative Analysis of YOLO Algorithm (Review)*', *IJECE*, Vol. 16, No. 1, p. 450, Feb. 2026, DOI: 10.11591/ijece.v16i1.pp450-462.

[31] M. A. A. Adam and J. R. Tapamo, '*Enhancing YOLOv5 for Autonomous Driving: Efficient Attention-based Object Detection on Edge Devices*', *J. Imaging*, Vol. 11, No. 8, p. 263, Aug. 2025, DOI: 10.3390/jimaging11080263.

[32] S. Potharaju, S. N. Tambe, K. Dasari, N. Srikanth, R. Venkatarao, and S. Tambe, *'Enhanced X-Ray Image Classification for Pneumonia Detection using Deep Learning based CBAM and SE Mechanisms*', *Intelligence-Based Medicine*, Vol. 12, p. 100299, 2025, DOI: 10.1016/j.ibmed.2025.100299.

[33] Z. Kang, Y. Liao, S. Du, H. Li, and Z. Li, '*SE-CBAM-YOLOv7: An Improved Lightweight Attention Mechanism-Based YOLOv7 for Real-Time Detection of Small Aircraft Targets in Microsatellite Remote Sensing Imaging*', *Aerospace*, Vol. 11, No. 8, p. 605, Jul. 2024, DOI: 10.3390/aerospace11080605.

 ISSN:2302-8149
e-ISSN:2540-9719

[34] T. M. M. Aung and A. A. Khan, '*Enhanced U-Net with Attention Mechanisms for Improved Feature Representation in Lung Nodule Segmentation*', *CMIR*, Vol. 21, p. e15734056386382, Oct. 2025, DOI: 10.2174/0115734056386382250902064757.

[35] N. Chandra, H. Vaidya, S. Sawant, S. Gite, and B. Pradhan, '*Attention Driven YOLOv5 Network for Enhanced Landslide Detection using Satellite Imagery of Complex Terrain*', *CMES*, Vol. 143, No. 3, pp. 3351–3375, 2025, DOI: 10.32604/cmes.2025.064395.

[36] Y. Mao and S. Pan, '*Steel Surface Defect Detection Method based on MAA_YOLOv8'*.

[37] S. Yang, J. Xing, Z. Liu, and X. Zheng, '*HCA-YOLOv8: A Pedestrian Detection Model that Integrates Multi-Source Context-Aware Mechanisms*', *IEEE Access*, Vol. 14, pp. 48487–48503, 2026, DOI: 10.1109/ACCESS.2026.3673396.

[38] S. Shao *et al.*, '*CrowdHuman: A Benchmark for Detecting Human in a Crowd*', Apr. 30, 2018, *arXiv*: arXiv:1805.00123. DOI: 10.48550/arXiv.1805.00123.

[39] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, '*WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild*', Sep. 25, 2019, *arXiv*: arXiv:1909.12118. DOI: 10.48550/arXiv.1909.12118.

[40] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, '*MOT16: A Benchmark for Multi-Object Tracking*', May 03, 2016, *arXiv*: arXiv:1603.00831. DOI: 10.48550/arXiv.1603.00831.

[41] P. Dendorfer *et al.*, 'MOT20: *A Benchmark for Multi Object Tracking in Crowded Scenes*', Mar. 19, 2020, *arXiv*: arXiv:2003.09003. DOI: 10.48550/arXiv.2003.09003.