IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2379-2387
Journal homepage: http://iieta.org/journals/ts

YOLO-OSAM: Reassembly Spatial Attention Mechanisms for Facial Expression Recognition

Ahmed Oday[1,2*], Azizi Abdullah[1], Shahnorbanun Sahran[1]

[1] Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, Bangi 43600, Malaysia
[2] Bioinformatics Department, College of Biomedical Informatics, University of Information Technology & Communications, Baghdad 10001, Iraq

Corresponding Author Email: p109064@siswa.ukm.edu.my

https://doi.org/10.18280/ts.420445

Received: 13 May 2025
Revised: 10 June 2025
Accepted: 18 June 2025
Available online: 14 August 2025

Keywords:
attention mechanism module, facial expression recognition, facial detection, YOLOv5

ABSTRACT

Facial Expression Recognition (FER) is crucial for accurately interpreting human emotions in human-computer interactions. However, FER remains challenging due to many variations, such as facial expressions, head poses, and illumination. Spatial attention mechanisms in Convolutional Neural Networks (CNNs) help address these challenges by enhancing feature extraction, directing focus to crucial facial regions while suppressing irrelevant information. However, traditional spatial attention modules, which apply average and max pooling followed by a convolutional layer, may have limited capacity to capture complex spatial dependencies, leading to suboptimal feature representation in FER tasks. YOLOv5 was selected from among various YOLO series because of its ability to deliver high accuracy object detection and classification, its lightweight architecture, and overall efficiency to overcome these limitations, we propose YOLO-OSAM, an enhanced version of YOLOv5 designed to refine feature learning for FER. Our approach introduces (1) a fusion layer that integrates attention mechanisms to ensure robust feature extraction across varying facial expressions and (2) an enhance spatial attention mechanism incorporated into the YOLOv5 architecture to capture fine-grained facial details. In this paper, the proposed attention mechanism module separately applies max and average pooling to generate feature maps, which are refined through three convolutional layers with batch normalization and Leaky ReLU activation. These processed maps are then concatenated and further optimized using additional convolutional blocks and SoftMax activation, with residual connections enhancing feature representation. Finally, we integrate this enhanced attention mechanism into the YOLOv5 neck, improving feature extraction and refinement. Experimental results demonstrate that YOLO-OSAM achieves 79.7%, 41.7%, and 98.1% accuracy on the RAF-DB (basic), RAF-DB (compound), and CK+ datasets, respectively—outperforming the original YOLOv5 by 1.3%, 0.8%, and 1.6%. Additionally, YOLO-OSAM surpasses baseline models such as VGG16, YOLOv3, and YOLOv5, highlighting its effectiveness in enhancing FER through improved spatial attention and feature extraction.

1. INTRODUCTION

Facial expressions are critical non-verbal communication methods that convey emotions and facilitate understanding during interactions. Consequently, some organizations and employers have adopted Facial Expression Recognition (FER) technology to analyze emotions, providing valuable insights into employees' emotional responses and enhancing their understanding and interactions [1]. Six primary facial emotions were identified to categorize facial movements into the following motor units: anger, disgust, fear, happiness, sadness, and surprise [2]. FER technology includes basic and advanced techniques, with the latter utilizing deep learning. Deep learning approaches outperform traditional methods by directly learning from input images, eliminating the need for costly pre-processing and feature extraction [3, 4]. This feature makes deep networks suitable for handling complex, large-input or output space problems, such as image and speech recognition. For instance, Convolutional Neural Networks (CNNs) [5] are well-suited for computer vision tasks [6]; Faster R-CNNs [7] and single-shot multi-box detectors [8, 9] are effective in object detection; VGG-Net [9] and GoogLeNet [10] are renowned architectures for image classification. These models have been pivotal in advancing computer vision and pattern recognition [11, 12]. Zhu et al. [13] used a CNN to reconstruct frontal-view images from canonical-view images based on the consistency and clarity of the face images, minimizing reconstruction loss error.

Kahou et al. [14] introduced You Only Look Once (YOLO), and YOLOv5 has outpaced YOLOv4 in real-world applications, object detection speed, owing to algorithm advancements. It provides four network configurations

(YOLOv5x, YOLOv5l, YOLOv5m, and YOLOv5s) to suit different object detection requirements by adjusting the width and depth of the feature extraction [15]. This led to the proposal of a network with an attention mechanism for automatic FER. The network comprises four components: feature extraction, attention, reconstruction, and classification. The attention mechanism lets the network focus on important features, improving the model's efficiency. Combining features with attention mechanisms can improve the attention models and yield better results. To enhance small-scale defect detection, a new fusion layer was introduced in YOLOv5 [16]—this layer integrates shallow features, critical for identifying subtle defects, and generates high-resolution feature maps. By emphasizing minor details and reducing background interference, the fusion layer improves localization precision and multi-scale detection. However, the method relies heavily on sufficient training data and struggles with misclassification when defect types have similar visual features, highlighting the need for further dataset augmentation and refinement. Zhu et al. [17] improved the YOLOv5 framework by incorporating a new fusion layer to address challenges in detecting small-scale from images. The original YOLOv5 struggled with capturing shallow features essential for detecting small objects. The new fusion layer generates high-resolution feature maps, enhancing the model's sensitivity to subtle features of small objects. By integrating shallow features from the backbone network into the fusion layers, the approach reduces feature loss and improves localization accuracy. However, the method has limitations, including its reliance on sufficient training data and a precision, which leaves room for improvement.

Attention mechanisms can identify salient regions and focus on features relevant to emotions, resulting in a more efficient representation of facial expressions [18]. Spatial attention mechanisms that can be embedded into end-to-end training and automatically determine the Region of Interest without manually cropping the image were developed by imitating human visual attention [19]. Woo et al. [20] lacked efficient feature selection, often overlooking relevant spatial and channel information. Conversely, the spatial attention mechanism allows the model to focus on key regions by redistributing weights across locations in the image. This process involves identifying critical areas to improve the prediction accuracy for those areas, such as objects or landmarks. Sun et al. [21] designed a shallow CNN and applied attention mapping to a fully connected layer. Marrero Fernandez et al. [22] created a dual-branch network that simultaneously extracts features and generates an attention map. However, traditional single-attention mechanisms used only for high-level features are inadequate for handling major pose, illumination, and occlusion variations. Thus, improving the emotional relevance of the features while controlling the model's complexity is crucial. Maximum (Max) and Average Pooling are used in deep neural networks, particularly in the spatial attention mechanism. Max Pooling selects the maximum values from each neighbouring region, allowing the extraction of the most important image features, such as edges or highlights, which helps improve the model's ability to focus on critical areas [23], whereas Average Pooling calculates the average values in each region, giving a general idea of the spatial distribution of signals. This helps to enhance the contextual information available in the image [24]. Max and Average Pooling treat each part of the image separately. Therefore, they may not adequately capture the contextual relationships between different parts, affecting the performance in tasks requiring an understanding of complex contexts. Pooling substantially reduces the data size, possibly leading to a loss of some structural information related to the shape or pattern in the image, particularly in the early stages of the network. Owing to the varying sizes of face images, recognizing facial expressions from such face images using object detection techniques, such as YOLO version, that use uniform-sized images as input presents challenges. Consequently, large faces may be segmented into different parts, causing the misidentification of a single expression as multiple expressions. Additionally, images returned by the proposed method often lack contrast, appearing grey and blurred, with faces blending into the background and aggregating, affecting the detection of smaller faces. Pretrained models, data augmentation, and post-processing techniques have been recommended to address these issues [25]. The Convolutional Block Attention Module (CBAM) addresses this limitation by enhancing essential features and suppressing irrelevant ones. CBAM is an advanced attention model used in deep neural networks to improve performance in image analysis and feature extraction. CBAM combines two attention mechanisms—channel and spatial. This combination allows the model to focus on the most important features and regions in an image. The channel attention mechanism compresses features across channels, generating weights for each channel based on its relevance to the input image. This enables the model to identify the crucial channels for each image, thereby improving prediction accuracy. This mechanism extracts information from channels and analyses their relative importance for a task [26]. Even with its lightweight features, CBAM faces the challenge of incurring computational overhead through channel and spatial attention modules, which potentially causes concern for real-world applications.

Despite the advancement of attention-based techniques such as CBAM for improving feature exploitation in deep neural networks, they are not yet used in practical one-shot detection systems for FER. Most available models either overfit to large-scale datasets or do not capture subtle facial features. Besides, many of these models were developed for object detection or classification, and are not optimized for emotion-specific spatial cues. This presents an opportunity to design a lightweight attention mechanism tailored for FER, especially in one-stage detectors like YOLOv5. To address this, we propose YOLO-OSAM, which augments spatial attention with a repetition-optimized CBL (Convolution-BatchNorm-LeakyReLU) module, providing efficient computation enhancements. This study enhances YOLOv5 detection and classification methods by incorporating attention mechanisms, which are commonly employed to extract more nuanced information from input objects. The input objects were images from the Real-world Affective Faces Database (RAF-DB), Cohn-Kanade (Ck+) dataset, Drowsiness dataset, and SC6-Net dataset. The proposed method adaptively locates the crucial regions and focuses on emotionally related features by introducing a spatial attention mechanism into the YOLOv5 model. We aimed to enable the detection and classification method to focus on the regions of interest in images, enhancing its performance in face detection. The main contributions of this study can be summarised as follows:

(1) An auxiliary branch is introduced into the feature-fusion layer to enhance multiscale facial detection, allowing for a more comprehensive capture of facial details.

(2) Attention modules are integrated into each linear layer's backbone to selectively amplify essential features, optimizing the information for facial emotion recognition before fusion.

(3) A spatial attention mechanism that applies a three-layer convolutional structure with residual connections to refined feature maps is developed, leveraging pooled features from Max and Average Pooling layers.

2. METHODS

FER is a fundamental component of human-computer interaction that allows systems to recognise and respond to human emotions. This section briefly introduces the YOLO detection method, which is based on deep learning and spatial attention mechanisms. Additionally, the improvements made to YOLOv5 and spatial attention mechanisms are discussed.

2.1 YOLOv5 method

The YOLOv5 architecture comprises three main sections: the backbone, neck, and head. In this study, CSPDarknet53 is used as the backbone for YOLOv5, and it incorporates the focus, CBL, CSP1, and SPP layers. The backbone—CSPDarknet53—consists of feature extraction layers such as convolutional, pooling, and other operations for feature map extraction and selection. It converts the input image into a set of feature maps that are then used by subsequent layers [27]. The architecture begins with the focus module, which divides the 4×4 input image into smaller 2×2 regions and independently extracts features from each region. The CBL module comprises convolution, batch normalization [28], and LeakyReLU [29] activation functions. Subsequently, the cross-stage partial network unit [30] was used. However, there were two types: CSP1_X in the backbone and CSP2_X in the neck network, where X represents the number of remaining units. An SPP module was added at the end of the backbone to enhance the architecture's performance, improving its ability to handle objects of different scales and aspect ratios. The second section in YOLOv5 is the neck layer, which combines the layer feature maps from different levels to minimize information loss. This involves the FPN [31] and PANet [32]. The FPN allows for the transmission of solid semantic features from higher to lower feature maps. Conversely, the PANet structure transmits robust localization features to move from lower to higher feature maps. The neck also contains a CSP2_X and the Concept module. The last section of the YOLOv5 is the head, which is responsible for making predictions, i.e., creating bounding box predictions and the associated class probabilities [33].

2.2 Spatial attention mechanism

A spatial attention module is a component of deep-learning models that improves their ability to focus on the most relevant parts of an image. This module helps the network concentrate on critical features by highlighting important spatial regions, thereby enhancing the performance of the model in FER tasks [34]. The spatial attention was calculated by taking the Average and Max Pooling values along the channel axis and combining them to create a concise feature descriptor [35, 36]. This method effectively identifies critical areas on the concatenated feature descriptor. A convolution layer was applied to generate a spatial attention map—$\mathrm{M_S} \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}}$. The attention map M_S assigns weights to different spatial regions, indicating the importance of each region. Regions with higher weights are considered more relevant for the task, and spatial attention is calculated as in Eq. (1) [18].

$$\mathrm{M_S} = \sigma\left(\mathrm{f}^{7\times7}([\mathrm{AvgPool(F)};\mathrm{MaxPool(F)}])\right) \tag{1}$$

where, F is the input feature map, σ represents the sigmoid activation function, and $f^{7\times7}$ represents a convolution operation with a 7×7 kernel.

2.3 Improve YOLOV5 and spatial attention

The YOLOv5s architecture was enhanced with new features to improve face detection and classification. First, a feature fusion layer was added to obtain additional information regarding the larger bounding box surrounding the face. This component captures more feature information for multiple face detection metrics, thereby enhancing the model's ability to extract features and make more accurate predictions. The features from the backbone network were derived from the first CSP1_1 unit, which is connected to the CBL unit. This unit is related to CSP1_3 and undergoes pooling, upsampling, and concatenation with CSP2_1, followed by a second convolutional unit to prevent information loss in more prominent faces. Second, attention modules were integrated into the backbone of each linear layer. These modules emphasize features critical for facial emotion classification, optimally preparing them for further processing in the neck layers. Figure 1 illustrates the improvement of YOLOv5 with the attention model, where the newly added modules are highlighted as rectangles.

2.3.1 Improve YOLOv5

Added a new fusion layer to the neck network to enhance YOLOv5s performance in detecting more prominent objects, generating a larger feature map with width/2 × height/2 × 3 dimensions. Combining the original network with fused feature maps achieved this enhancement, and they were then up-sampled and combined into four layers. These feature maps measured width/2 × height/2 × 3, width/4 × height/4 × 3, width/8 × height/8 × 3, and width/16 × height/16 × 3, corresponding to four scales derived from the backbone network. The network uses four channels, each with different feature map sizes enhancing its effectiveness. Thus, the head comprises four detection heads and locates and classifies the feature information output from the neck, providing classification probabilities, confidence scores, bounding boxes, and other relevant information for each detected target. LeakyReLU enables the network to learn intricate patterns and relationships in the data by maintaining a slightly positive slope for negative inputs. LeakyReLU enables gradient flow through negative inputs; nonetheless, it must maintain uniformity during training. Adjusting the learning rate and periodically evaluating the neural network can help identify the optimal configuration ratio for this activation function to address this challenge, as indicated in Eq. (2). Batch normalisation enhances the back-propagation of gradients, prevents gradient attenuation, and preserves emotional feature information, as defined in Eq. (3). Inspired by residual networks, where (x) denotes the input, μ(mu) indicates the mini-batch mean, σ(sigma) represents the standard deviation (a measure of the variance in the input), and γ(gamma) and β(beta) are parameters for nonormalizingnd shifting values, respectively.

$$\text{Leaky ReLU}\ (x) = \begin{cases} x & \text{if } x > 0 \\ ax & \text{if } x \leq 0 \end{cases} \quad (2)$$

where, σx is a small positive constant.

$$\text{Bn}\ (x) = \gamma(\sigma x - \mu) + \beta \quad (3)$$

2.3.2 Improve spatial attention

Attention modules were added to the output of the backbone layers to enhance facial information in the images. Specifically, spatial attention modules were implemented to enhance the FER accuracy. These modules assign higher weights to emotion-rich facial regions and iteratively partition the face into blocks, enabling targeted processing for FER tasks. This is particularly beneficial in complex models, where the spatial attention module can adapt to changing backgrounds, providing a more robust deep learning network across diverse environments. Without this module, the feature extraction module may capture irrelevant information from the entire image, resulting in less representative information. The spatial attention mechanism was improved by applying a 3 × 3 CBL module layer (repeated thrice) to the outputs of the Max and Average Pooling layers to produce a focused facial attention map corresponding to the Region of Interest. This was achieved by connecting channels and combining elements generated from the feature map to simulate a residual block structure. The final feature map was obtained by applying a SoftMax activation function to the combined feature map. Figure 2 illustrates the proposed spatial attention module. It takes a feature map (F) and does average-pooling and max-pooling on it in parallel, the pooling method captures spatial context. The pooled features are concatenated and refined through several CBL (Convolutional, BatchNorm, Leaky ReLU) layers to increase the accuracy of the spatial information. A SoftMax function, which generates a spatial attention map, is also included in the model along with the feature map. These two models operate together to highlight important areas within the feature map. This module improves the model's ability to focus on relevant spatial details, enhancing performance in tasks like object detection or image segmentation.

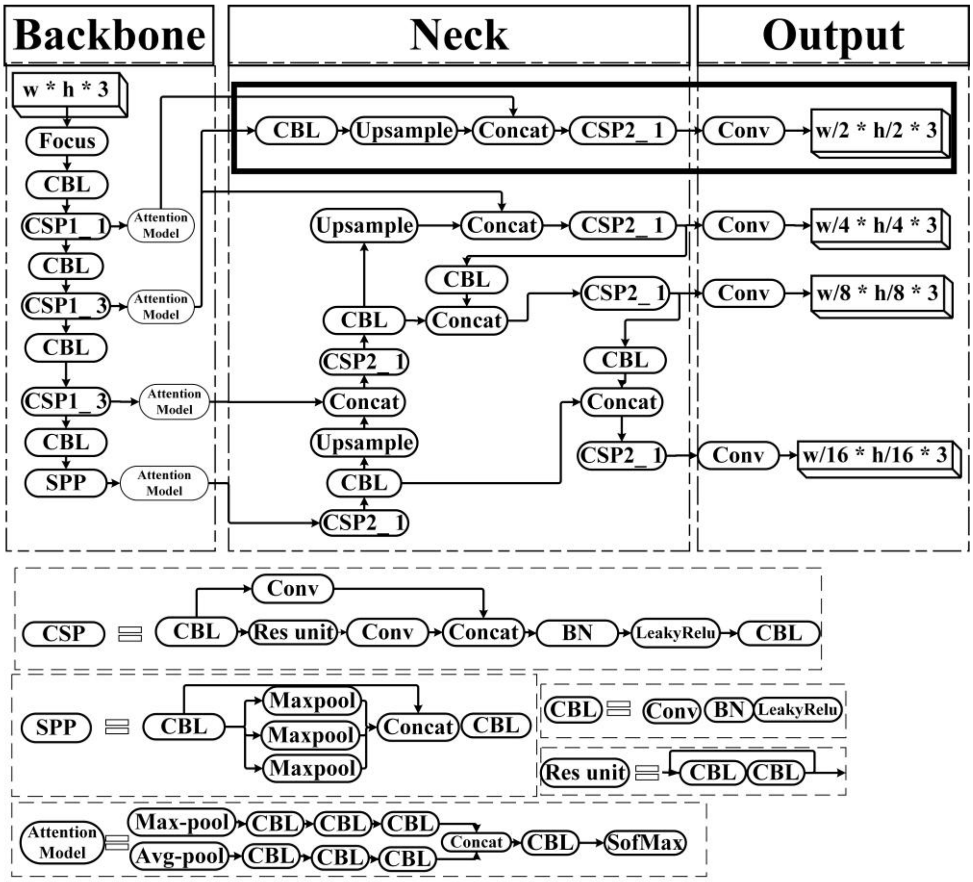

Figure 1. YOLO-OSAM model architecture for the improvement of YOLOv5s with the attention model

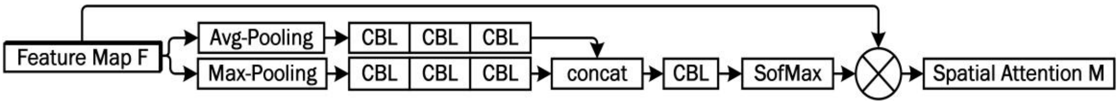

Figure 2. The proposed spatial attention mechanism, a model responsible for achieving optimal features

The input is an intermediate feature map, denoted as $F \in R^{C\times H\times W}$, with dimensions C×H×W. C is the number of channels, and H and W are the height and width of the feature map. The spatial attention module applies Average and Max Pooling to obtain two independent 2D maps, which are then processed thrice by 3×3 CBL layers to further expand the receptive fields and effectively use the contextual information of each map:

$$\mathrm{Mi} = \left(3 \times \mathrm{CBL}\left(\mathrm{MAX\ Pooling\ (F)}\right)\right) \tag{4}$$

$$\mathrm{Ni} = \left(3\mathrm{CBL}\left(\mathrm{AveargePooling(F)}\right)\right) \tag{5}$$

Next, the outputs of the 2D spatial attention map M_i, N_i are concatenated to obtain attention map Si, which is subsequently input into a CBL module and evaluated by a SoftMax activation function to generate attention maps A:

$$\mathrm{Si} = \mathrm{concat}\ (\mathrm{Mi}, \mathrm{Ni}) \tag{6}$$

$$\mathrm{A} = \mathrm{Softmax}\left(\mathrm{CBL(Si)}\right) \tag{7}$$

Finally, map A is multiplied by the input feature map F to enhance the representability of the feature map, resulting in the channel attention map $\mathrm{T_{SPatial}} \in \mathrm{R^{C\times1\times1}}$.

$$\mathrm{T_{spatial}} = \mathrm{A} \times \mathrm{F} \tag{8}$$

Thus, the spatial attention module can attract different types of attention and identify important features more accurately by focusing on local and global perceptions.

3. DATASET

This section describes the application of two datasets to the improved YOLO-OSAM method: RAF-DB (basic and compound) of real facial expressions and CK+ laboratory facial expressions. These datasets are used to train and evaluate deep learning models, with 70% of the data used for training, 10% for validation, and 20% for testing, Although the 70-10-20 data split is not the most conventional choice, it has been successfully employed in several prior studies in the fields of emotion recognition and object detection [2]. In our case, this configuration allowed for a relatively large test set to ensure reliable evaluation, while preserving enough data for training and validation for comparing the results.

This CK+ database [37, 38] It is the smallest among the laboratory datasets. The image data represent the facial expressions of individuals aged 18–50 years, with a distribution of 69% female and 31% male participants from various nationalities. The dataset includes seven facial expressions: surprise: 249, fear: 75, disgust: 177, happiness: 207, sadness: 84, anger: 135, and contempt: 54.

The RAF-DB-basic dataset initially consisted of 15,339 facial images collected from real scenes. This extensive dataset offers excellent generalizability and robustness. It includes single-expression labels: surprise: 1619, fear: 355, disgust: 877, happiness: 5957, sadness: 2460, anger: 867, and neutral: 3204.

The RAF-DB-compound dataset [39] is derived from the RAF-DB basic scenes contains 3,954 facial images, The distribution of these emotions is: happily surprised: 697, happily disgusted: 266, sadly fearful: 129, sadly angry: 163, sadly surprised: 86, sadly disgusted: 738, fearfully angry: 150, fearfully surprised: 560, angrily surprised: 176, angrily disgusted: 841, and disgusted and surprised: 148.

The FER2013 [40] is an emotion category through crowdsourcing that builds upon. The quality of emotion-recognition tasks is improved by correcting the labelling errors. It consists of 35,887 facial images grouped into eight emotions: anger, disgust, fear, happiness, neutrality, sadness, surprise, and contempt.

3.1 Experiment

YOLO-OSAM experiments evaluated the FER model by comparing its accuracy and mean Average Precision (mAP) using a score threshold of 0.5. Accuracy is the percentage of correct predictions, which is determined by dividing the number of correctly classified samples by the total number of samples (Eq. (9)). Regarding precision, the accuracy of positive predictions was calculated as the ratio of correctly predicted positive instances to the total number of true positives and false positives (Eq. (10)). In addition, recall, sometimes called the sensitivity or true-positive rate, assesses a model’s capacity to accurately identify pertinent instances. It is determined by dividing the number of true positives (correctly forecasted positive cases) by the sum of true positives and false negatives, where false negatives represent positive cases incorrectly classified as negative (Eq. (11)) [41]. Moreover, we calculated the mAP across all classes. Thus, average precision, which has evolved, is not equal to the average Precision (P), as expressed in Equation 12. Therefore, the average precision can be understood as the area under the precision-recall curve. mAP@0.5 refers to the mAP at the IoU threshold of 0.5. Early stopping for the optimal model halts the training process when errors in the validation dataset increase, preventing overfitting, which would otherwise be evident in the training dataset.

$$\mathrm{Accuracy} = \frac{\mathrm{n\ correct}}{\mathrm{n_{total}}} \tag{9}$$

where, *n* correct is the number of correctly classified samples.

$$\mathrm{Precision\ (P)} = \frac{\mathrm{True\ Postive\ (TP)}}{\mathrm{True\ Postive\ (TP)+False\ Postive\ (FP)}} \tag{10}$$

$$\mathrm{Recall\ (R)} = \frac{\mathrm{True\ Postive\ (TP)}}{\mathrm{True\ Postive\ (TP)+False\ Negtive\ (FN)}} \tag{11}$$

$$\mathrm{mAP} = \frac{1}{\mathrm{N_c}}\sum_{\mathrm{t=i}}^{\mathrm{N}} \mathrm{P(t)\Delta R(t)} \tag{12}$$

where, *P*, *R*, and *Nc* represent the precision, recall rate, and number of classes, respectively.

4. RESULTS AND DISCUSSION

4.1 Training setting

The NVIDIA GeForce 4080 RTX GPU with 12GB of memory and 64GB of computer memory was used for training. The Stochastic Gradient Descent (SGD) optimiser was applied to RAF-DB-basic and RAF-DB-compound. However, the Adam optimiser was applied to the CK+ dataset. The momentum was set at 0.937, with a batch size of 32 and an initial learning rate of 0.001. The model was trained for 500 epochs using early stopping to determine the optimal number. The code is publicly available at https://github.com/Ahmed-Oday/YOLOv5-Spatial-Attention-Mechanism.git.

4.2 Comparison of experiment results

Table 1 compares the performances of VGG-16, YOLO series, and the YOLO-OSAM proposed method across the RAF-DB-basic, RAF-DB-compound, CK+, and FER2013 datasets. These results highlight the robustness and effectiveness of the YOLO-OSAM method, particularly on complex datasets, where it consistently ranked as a top performer across all evaluated metrics.

Table 1. Comparison with related methods

Model	Dataset	F1-Score(%)	mAP0.5 (%)	Accuracy (%)	Opt. Epoch
VGG-16	RAF-DB-basic	74.08	78.50	70 [42]	63
	RAF-DB-compound	28.04	30	31	76
	CK+	95.98	96.20	95.10 [43]	40
	FER2013	64.9	67.3	65.2	73
YOLOv3	RAF-DB-basic	69.66	73 [24]	73.10	101
	RAF-DB-compound	36.71	33	30	120
	CK+	80.57	96.80	99 [24]	94
	FER2013	69.35	68.4	70.3	112
YOLOv5	RAF-DB-basic	80.70	86.10	78.40	278
	RAF-DB-compound	45.04	46.20	40.90	298
	CK+	93.73	99.4	96.50	500
	FER2013	73.05	70.13	72.24	137
YOLOv6	RAF-DB-basic	79.14	79.20	78.20	146
	RAF-DB-compound	37.14	40.10	39.20	188
	CK+	94.57	96.10	96.40	144
	FER2013	70.85	69.5	68.4	116
YOLOv7	RAF-DB-basic	77.73	81.20	79.20	183
	RAF-DB-compound	36.43	40.10	39.10	201
	CK+	95.03	97.10	96.60	198
	FER2013	73.6	71.2	70	
YOLOv8	RAF-DB-basic	80.78	83.10	78.60	286
	RAF-DB-compound	41.82	43.20	40.20	173
	CK+	94.67	97.20	97.60	185
	FER2013	74.6	72.5	73.6	169
YOLOv9	RAF-DB-basic	79.18	82.10	77.20	215
	RAF-DB-compound	40.33	40.40	40.20	200
	CK+	93.72	98.50	97.60	206
	FER2013	73.8	71.7	72.5	157
YOLOv10	RAF-DB-basic	80.57	84.10	78.60	149
	RAF-DB-compound	36.57	40.90	39.70	100
	CK+	94.01%	94.20	94.90	164
	FER2013	72.5	69.7	73.1	140
YOLOv11	RAF-DB-basic	80.67%	84.80%	79.10	267
	RAF-DB-compound	43.17%	44.20%	40.40	228
	CK+	96.69%	98.70%	97.90	211
	FER2013	74.3	71.5	73.7	180
YOLOv12	RAF-DB-basic	80.10%	82.30%	78.90	193
	RAF-DB-compound	37.64%	35.20%	38	165
	CK+	95.88%	96.90%	96.80	201
	FER2013	71.8	74.7	72.5	143
YOLO-OSAM (propose)	RAF-DB-basic	81.26%	85.70%	79.70	347
	RAF-DB-compound	45.70%	45.90%	41.70	301
	CK+	97.64%	99.50%	98.10	483
	FER2013	78.2	76.4	75.4	217

4.3 Effectiveness of the proposed spatial attention on the CK+ dataset

The evaluation of different spatial attention mechanisms on the CK+ datasets focused on metrics such as precision, recall, mAP50 (mAP at 50% IoU), and accuracy to assess the effectiveness of each method in detecting and classifying facial expressions accurately. in the CK+ dataset, Table 2 compares the performance of YOLOv5 models with standard spatial attention, the proposed spatial attention, and the baseline of CSPDarknet-53, with and without the spatial attention mechanism, across different versions of CNN. The proposed spatial attention restored perfect precision and recall scores, demonstrating its effectiveness in enhancing the model's performance. To assess the statistical significance of the performance improvements achieved by the proposed YOLO-OSAM module, independent samples t-tests were conducted across the CK+ datasets. The difference is statistically significant in Table 3.

4.4 Analysis of experimental results CK+ dataset

4.4.1 Comparison of CBL modules in the spatial attention mechanism in different models

We conducted experiments to study the effect of the repeated number of CBL (CBL=0, 1, 2, 3, and 4) on OSAM in three models based on the CK training dataset. Table 4 presents the analysis of the three models—CNN, CSPDarknet-

53 (baseline), and YOLOv5—across different configurations of the CBL parameter. In addition, it provides the average performance of these models and analyses the stability of the CBL values (0, 1, 2, 3, and 4). This indicates consistent performance with minimal variation. However, CBL of 0 and 4 had greater variability, suggesting lower stability.

4.4.2 Experiments on spatial attention mechanism at different locations

We demonstrated the importance of the location of OSAM in the model. Two models were assessed: the CSPDarknet53 model and its variant. The results are presented in Table 5, where CSPDarknet53-embedding indicates the location of the spatial attention mechanism applied after each CSP1 unit in CSPDarknet53 (backbone). CSPDarknet53-last represents the spatial attention mechanism applied after SPP at the end of CSPDarknet53 (backbone). The comparison reveals that CSPDarknet53-last outperforms CSPDarknet53-imbedding in accuracy.

The second model is YOLOv5, where YOLOv5-Neck represents the location of OSAM at the end of the neck, specifically applied to every branch from the neck to the head unit after CSP2_1. YOLO-OSAM, in contrast, applies the spatial attention mechanism at the end of the backbone, implying that it was applied from every branch from the backbone to the neck unit Figure 1. Table 6 presents the results, revealing that YOLO-OSAM outperforms YOLOv5-Neck.

Table 2. Effectiveness in the Ck+ dataset

Model	F1-Score (%)	Map (%)	Accuracy (%)
CSPDarkNet53	79.43	91.2	76.8
CSPDarkNet53 + Spatial AM [18]	65.02	85.9	72.4
CSPDarkNet53 + OSAM	69.74	85.9	81.7
CSPDarkNet53 + CBAM (Channel AM + OSAM)	96.58	98	91.7
CSPDarkNet53 + CBAM (Channel AM + Spatial AM) [20]	95.47	98.4	91.7
AlexNet [44]	N/A	N/A	98.38
AlexNet [44] + Spatial AM [18]	97.64	97.7	97.4
AlexNet [44] + OSAM	98.4	98.4	98.4
AlexNet [44] + CBAM (Channel AM + OSAM)	98.4	98.4	98.4
AlexNet [44]+ CBAM (Channel AM + Spatial AM) [20]	96.64	96.9	96.4
Detection			
YOLOv5 [17] (w/o adding a layer)	93.73	99.4	96.5
YOLOv5 [17] + Spatial AM [18]	95.46	99.1	96.5
YOLOv5 [17] + OSAM	97.59	99.5	96.1
YOLOv5 [17]+ CBAM (Channel AM + OSAM)	91.74	99.2	97.5
YOLOv5 [17] + CBAM (Channel AM + Spatial AM) [20]	96.34	99.2	93.1
Improve YOLOv5 (adding a layer)	97.09	99.4	95.2
Improve YOLOv5 + Spatial AM [18]	87.90	97.3	90.5
Improve YOLOv5 + OSAM	97.64	99.5	98.1
Improve YOLOv5 + CBAM (Channel AM + OSAM) [20]	98.73	99.5	99.5
Improve YOLOv5+ CBAM (Channel AM + Spatial AM) [20]	88.85	96.2	88.4

Table 3. Statistical comparison of baseline models vs. YOLO-OSAM models using an independent samples t-test

Dataset	Round	Baseline CSPdarknet-53		YOLO-OSAM		T-Value	P-Value
		Mean (%)	STD (%)	Mean (%)	STD (%)		
CK+	10	76.8	8.26	86.33	6.84	2.81	0.01

Table 4. Most stable CBL analysis

Model	CBL=0	CBL=1	CBL=2	CBL=3	CBL=4
AlexNet [44]	97.8	98.2	97.9	**98.4**	95.9
CSPDarknet-53	98.1	94.1	91.7	**98.2**	94.4
YOLOv5 [17]	90.1	96.4	**96.5**	95.7	95.7

Table 5. Comparison of the location of the spatial attention mechanism in CSPDARKNET53(train dataset)

CSPDarknet53–CK-Location	Accuracy	mAP
CSPDarknet53-embedding	88%	96.7%
CSPDarknet53-last	93%	97.4%

Table 6. Comparison of the location of the spatial attention mechanism in YOLOv5 (train dataset)

YOLOV5–CK-Location	Accuracy	mAP
Improve-YOLOV5-Neck	0.627	0.893
Improve-YOLO-OSAM	0.981	0.992

5. CONCLUSIONS

This study improves a real-world FER system for computer interaction and vision, using an improved YOLOv5 with an attention mechanism. First, a new fusion layer was added to YOLOv5, generating YOLO-OSAM predictive heads. These heads minimised the effects of face size variations and enhanced the detection of small faces. Second, spatial attention units were added to each output branch of the backbone to highlight the information that contributed to the extraction of more features. Finally, the performances of deep learning models were compared, and the effects of the spatial attention mechanism and its location were evaluated using the RAF-DB-basic, RAF-DB-compound, and CK+ datasets.

Regarding the RAF-DB-basic dataset, the spatial attention mechanism slightly improved accuracy (79.7%) compared with standard spatial attention, indicating minor advancements in feature extraction and classification accuracy. On the more complex RAF-DB-compound dataset, spatial attention mechanisms performed similarly in terms of the mAP50 and accuracy metrics, suggesting that the YOLO-OSAM mechanism is beneficial for more challenging tasks. Furthermore, the CK+ dataset results highlighted a significant advantage of YOLO-OSAM, with a substantial increase in accuracy (98.1%) compared to the standard approach (96.5%), demonstrating its superior ability to handle datasets with clear and distinct features. These findings suggest that the custom spatial attention mechanism offers considerable benefits for tasks with well-defined features; however, its impact on more complex datasets is less pronounced, suggesting the need for further optimisation or alternative methods to tackle such challenging tasks.

ACKNOWLEDGMENT

This work has been supported by the Ministry of Higher Education Malaysia (Grant No.: FRGS/1/2024/ICT02/UKM/02/5) and the University Kebangsaan Malaysia Research University (Grant No.: GUP-2021-063).

REFERENCES

[1] Zhong, H., Han, T., Xia, W., Tian, Y., Wu, L. (2023). Research on real-time teachers' facial expression recognition based on YOLOv5 and attention mechanisms. EURASIP Journal on Advances in Signal Processing, 2023(1): 55. https://doi.org/10.1186/S13634-023-01019-W

[2] Zhi, J., Song, T., Yu, K., Yuan, F., Wang, H., Hu, G., Yang, H. (2022). Multi-attention module for dynamic facial emotion recognition. Information, 13(5): 207. https://doi.org/10.3390/INFO13050207

[3] Caroppo, A., Leone, A., Siciliano, P. (2020). Comparison between deep learning models and traditional machine learning approaches for facial expression recognition in ageing adults. Journal of Computer Science and Technology, 35: 1127-1146. https://doi.org/10.1007/s11390-020-9665-4

[4] Ali, G., Ali, A., Ali, F., Draz, U., Majeed, F., Yasin, S., Ali, T., Haider, N. (2020). Artificial neural network based ensemble approach for multicultural facial expressions analysis. IEEE Access, 8: 134950-134963. https://doi.org/10.1109/ACCESS.2020.3009908

[5] Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. Insights Into Imaging, 9: 611-629. https://doi.org/10.1007/S13244-018-0639-9

[6] Yan, J. (2024). Application of CNN in computer vision. Applied and Computational Engineering, 30: 104-110. https://doi.org/10.54254/2755-2721/30/20230081

[7] Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D. A., Deng, J., Sukthankar, R. (2018). Rethinking the faster R-CNN architecture for temporal action localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, pp. 1130-1139. https://doi.org/10.1109/CVPR.2018.00124

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, Netherlands, pp. 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[9] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), State, USA, pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[11] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3523-3542. https://doi.org/10.1109/TPAMI.2021.3059968

[12] Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z. (2020). Attention mechanism-based CNN for facial expression recognition. Neurocomputing, 411: 340-350. https://doi.org/10.1016/J.NEUCOM.2020.06.014

[13] Zhu, Z., Luo, P., Wang, X., Tang, X. (2014). Recover canonical-view faces in the wild with deep neural networks. arXiv preprint arXiv:1404.3543. https://doi.org/10.48550/arXiv.1404.3543

[14] Kahou, S.E., Michalski, V., Konda, K., Memisevic, R., Pal, C. (2015). Recurrent neural networks for emotion recognition in video. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 467-474. https://doi.org/10.1145/2818346.2830596

[15] Irsan, M., Hassan, R., Hasan, M.K., Lam, M.C. (2024). Exploring the power of convolutional neural networks in face detection. In Intelligent Systems of Computing and Informatics, New York, United States, pp. 98-113. https://doi.org/10.1201/9781003400387-7

[16] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91

[17] Zhu, L., Geng, X., Li, Z., Liu, C. (2021). Improving YOLOv5 with attention mechanism for detecting boulders from planetary images. Remote Sensing, 13(18): 3776. https://doi.org/10.3390/RS13183776

[18] Luh, G.C., Wu, H.B., Yong, Y.T., Lai, Y.J., Chen, Y.H. (2019). Facial expression based emotion recognition employing YOLOv3 deep neural networks. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, pp. 1-7. https://doi.org/10.1109/ICMLC48188.2019.8949236

[19] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. https://arxiv.org/abs/2004.10934v1

[20] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[21] Sun, W., Zhao, H., Jin, Z. (2018). A visual attention

based ROI detection method for facial expression recognition. Neurocomputing, 296: 12-22. https://doi.org/10.1016/J.NEUCOM.2018.03.034

[22] Marrero Fernandez, P.D., Guerrero Pena, F.A., Ren, T., Cunha, A. (2019). Feratt: Facial expression recognition with attention net. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, USA, pp. 837-846. https://doi.org/10.1109/CVPRW.2019.00112

[23] Zhang, K., Huang, Y., Du, Y., Wang, L. (2017). Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Transactions on Image Processing, 26(9): 4193-4203. https://doi.org/10.1109/TIP.2017.2689999

[24] Ma, H., Celik, T., Li, H. (2021). Fer-yolo: Detection and classification based on facial expressions. In International Conference on Image and Graphics, pp. 28-39. https://doi.org/10.1007/978-3-030-87355-4_3

[25] Li, Z., Song, J., Qiao, K., Li, C., Zhang, Y., Li, Z. (2022). Research on efficient feature extraction: Improving YOLOv5 backbone for facial expression detection in live streaming scenes. Frontiers in Computational Neuroscience, 16: 980063. https://doi.org/10.3389/FNCOM.2022.980063

[26] Wang, Y., Wang, W., Li, Y., Jia, Y.D., Xu, Y., Ling, Y., Ma, J.Q. (2024). An attention mechanism module with spatial perception and channel information interaction. Complex Intelligent Systems, 10: 5427-5444. https://doi.org/10.1007/s40747-024-01445-9

[27] Pann, V., Lee, H.J. (2022). Effective attention-based mechanism for masked face recognition. Applied Sciences, 12(11): 5590. https://doi.org/10.3390/app12115590

[28] Hou, Q., Zhou, D., Feng, J. (2021). Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, pp. 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350

[29] Su, Z., Adam, A., Nasrudin, M.F., Prabuwono, A.S. (2024). Proposal-free fully convolutional network: Object detection based on a box map. Sensors, 24(11): 3529. https://doi.org/10.3390/S24113529

[30] Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., Li, X. (2021). A real-time detection algorithm for Kiwifruit defects based on YOLOv5. Electronics, 10(14): 1711. https://doi.org/10.3390/ELECTRONICS10141711

[31] Thakkar, V., Tewary, S., Chakraborty, C. (2018). Batch normalization in Convolutional Neural Networks—A comparative study with CIFAR-10 data. In 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), Kolkata, India, pp. 1-5. https://doi.org/10.1109/EAIT.2018.8470438

[32] Mastromichalakis, S. (2020). ALReLU: A different approach on Leaky ReLU activation function to improve neural networks performance. arXiv preprint arXiv:2012.07564. https://arxiv.org/abs/2012.07564v2

[33] Kim, D., Park, S., Kang, D., Paik, J. (2019). Improved center and scale prediction-based pedestrian detection using convolutional block. In 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, pp. 418-419. https://doi.org/10.1109/ICCE-BERLIN47944.2019.8966154

[34] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, pp. 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[35] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, pp. 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

[36] Redmon, J., Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, pp. 7263-7271. https://doi.org/10.1109/CVPR.2017.690.

[37] Shi, G., Huang, J., Zhang, J., Tan, G., Sang, G. (2021). Combined channel and spatial attention for YOLOv5 during target detection. In 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), Chengdu, China, pp. 78-85. https://doi.org/10.1109/PRML52754.2021.9520728

[38] Gan, C., Xiao, J., Wang, Z., Zhang, Z., Zhu, Q. (2022). Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention. Image and Vision Computing, 117: 104342. https://doi.org/10.1016/J.IMAVIS.2021.104342

[39] Li, S., Deng, W., Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, pp. 2852-2861. https://doi.org/10.1109/CVPR.2017.277

[40] Khaireddin, Y., Chen, Z. (2025). Facial emotion recognition: State of the art performance on FER2013. arXiv preprint arXiv:2507.12345. https://doi.org/10.48550/arXiv.2507.12345.

[41] Almansour, N., Albashish, D., Sahran, S., Abdullah, A., Nasruddin, M.F., Xinlan, X. (2025). Metaheuristic-based hyperparameter tuning for pretrained deep learning model: Application to the skin cancer identification. In Proceedings of the 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA), Amman, Jordan, pp. 1-8. https://doi.org/10.1109/ICCIAA65327.2025.11013496

[42] Tutuianu, G.I., Liu, Y., Alamäki, A., Kauttonen, J. (2024). Benchmarking deep facial expression recognition: An extensive protocol with balanced dataset in the wild. Engineering Applications of Artificial Intelligence, 136: 108983. https://arxiv.org/abs/2311.02910v1

[43] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, USA, pp. 94-101. https://doi.org/10.1109/CVPRW.2010.5543262

[44] Aza, M.F.U., Suciati, N., Hidayati, S.C. (2020). Performance study of facial expression recognition using convolutional neural network. In 2020 6th International Conference on Science in Information Technology (ICSITech), Palu, Indonesia, pp. 121-126. https://doi.org/10.1109/ICSITech49800.2020.9392070