Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3383-3390
Journal homepage: http://iieta.org/journals/ts

YOLO-ISAM: Improved Spatial Attention Mechanism Model for Masked Face Detection

Shahad Fadhil[1*], Shaimaa Hameed Shaker[1], Firas A. Abdullatif[2]

[1] College of Computer Science, University of Technology-Iraq, Baghdad 10001, Iraq
[2] College of Education for Pure Science, Ibn-Al-Haithem, Baghdad University, Baghdad 10001, Iraq

Corresponding Author Email: cs.20.30@grad.uotechnology.edu.iq

https://doi.org/10.18280/ts.420626

Received: 10 September 2025
Revised: 12 October 2025
Accepted: 19 November 2025
Available online: 31 December 2025

Keywords:
face mask, improve-YOLO, masked face detection, spatial attention mechanism

ABSTRACT

Detailed Face masks present a significant challenge part, such as balancing privacy rights and public safety. Which rely on visible features and struggle with small faces, lighting, camera angles, to address this problem. This paper introduces YOLO-ISAM, an improved model to detect Masked Faces Recognition (MFR). Our approach network was upgraded with an Improved Spatial Attention Mechanism (ISAM) in the backbone and a new fusion layer of YOLO, which increasing the prediction contributions to face detection. The ISAM Module leverages parallel max, average, and median pooling operations to generate a more robust spatial attention map, allowing the model to focus on the most relevant facial regions obscured by masks. We evaluate the proposed model on the four dataset The ChokePoint, VIDMASK, Moxa3k, and LFW-SM datasets were used as benchmarks owing to the lack of video datasets with masked individuals. Experimental results demonstrate that YOLO-ISAM significantly outperforms of the masked face detection performance across multiple datasets, achieving Accuracy rates of 92%, 99%, 91%, and 100% on the respective datasets.

1. INTRODUCTION

Recent studies have proven their ability to accurately detect people with occluded faces, that is, whether they have occlusions (in the upper or lower part of the face) under low-light conditions [1]. However, the accuracy of the results is related to the database used, noting that all research that dealt with knowing the identity of a person who suffers from facial occlusion was based on data in which the person's face is very close [2, 3]. This is a challenge in this research, which is the person wearing the mask and tracking him from a distance (several meters), as is the case in offices, governmental and private institutions, and hospitals. Therefore, there is a need for further development to keep pace with these needs and changes in the real world [4, 5]. Deep learning-based categorization of object detection algorithms is into two groups: two-stage and one-stage approaches [6]. Two-stage methods, such as Region-based Convolutional Neural Networks (R-CNN), Fast R-CNN, Faster R-CNN, and Mask R-CNN, involve a two-step process where potential Regions of Interest (ROIs) are proposed in the input image and then classified and regressed. bounding boxes and class labels is one-stage methods that directly predict performing regression on a dense sampling of predefined locations in the image without first proposing regions [7]. The difference between them is that two-stage detection algorithms extract features and then determine the location of the target objects among the candidate regions. The one-stage method directly determines and classifies the location of an object [8]. A single-shot multi_box detector (SSD) [9] is an object-detection approach that relies on a single neural network. One of the methods for single-stage detection is the YOLO [10] series and SSD series; the YOLO series is faster but less accurate than the two-stage detectors. It is faster, that is, less time-consuming, by reducing the calculations that result from skipping the step of creating suggestion boxes, every box must detect and classify objects in an area, because of 7×7 square images divided [11]. YOLOv1 was less flexible as it required changing the size of the input because the last two layers in it were fully connected, hence the updates to it, as YOLOv2 [12] and YOLOv3 [13], An anchor mechanism was used to address these flaws it improves accuracy. YOLOv2 also uses the k-means algorithm to size the input to fit the width to height of the anchor boxes. The release of YOLOv4 represents many improvements compared to previous versions, such as mosaic data augmentation and integration of partial cross-stage connections (CSP) and merging them with Darknet53 to create CSPDarknet53, which in turn significantly reduces the computational complexity [14, 15]. Among these, YOLOv5 is lightweight and portable YOLOv5 also uses CSPDarknet53 to capture depth features from input images [16]. This creates a clear research gap: the lack of a robust detection model specifically optimized for the small-size and high-occlusion characteristics of masked faces at a distance.

A solution was presented to address the challenge of achieving an accurate mask detection under low-light conditions [17]. The proposed approach combines an attention mechanism with a CSPDarknet53 network model, and attention mechanisms have attracted significant interest in the field of deep learning. The Convolutional Block Attention Module "CBAM" combines both the Channel Attention Module (CAM) and Spatial Attention Module (SAM) to enrich the original feature map [18]. The primary weakness of this strategy is its generality and post-hoc application. Modules like CBAM are designed for general vision tasks and are not optimized for the specific spatial configuration of a masked face, where the upper half (eyes, brow) becomes critically important. Furthermore, simply adding these modules does not guide the backbone to fundamentally learn more robust features for occlusion; it only refines the features after they have been extracted. This often leads to suboptimal feature representation for the specific MFR task. Wang et al. [19] integrated attention modules. The Convolutional Block Attention Module (CBAM) has been a popular choice, with studies like reference [3] appending it to a backbone network to refine features. However, this post-hoc application of generic attention is often suboptimal, as it does not guide the feature extraction process from within the network's core layers. The spatial attention mechanism aims to highlight significant spatial features in the original feature map [18]. Researchers have investigated the impact of factors such as maximum pooling and average pooling within the channel attention module, as well as the effect of the order in which the channel attention module and SAM are applied on the performance of the model [20]. Spatial Attention is one of the most important aspects of neural networks, and it helps identify the most important parts of the input data; therefore, less important details can be neglected. Such a mechanism facilitates the increase in representations in the network by promoting more essential output areas to be selectively boosted [21, 22].

However, the quality of this imaging technique is sometimes poor, this may be affected by the use of different angles of the face, lighting conditions, partial or total occlusion, low resolution, or noisy imaging. For instance, researchers can eliminate certain factors (such as the darkening of face images or the less ideal positioning angle) by preprocessing face images and then selecting the most suitable angle for the hardware [23, 24].

This paper improves YOLO-ISAM for these challenges in detection methods by Improving Spatial Attention Mechanism (ISAM) approach. which we integrate directly into the CSPDarknet53 backbone. Unlike simply adding attention modules, our ISAM is structurally embedded, using parallel max, average, and median pooling paths to generate a more powerful spatial attention map. This allows the network to dynamically focus on the most salient and visible facial regions—such as the eyes and forehead—while suppressing irrelevant background clutter, which is crucial for detecting small, occluded faces. In our experiments, the input data consisted of three datasets: ChokePoint, VIDMASK, Moxa3k, and LFW-SM. By integrating the attention mechanism into the YOLO-ISAM model, our objective is to refine the detection process, enabling the model to focus on key regions within the images, thereby improving its effectiveness in detecting masked faces. Simultaneously, the improved model aims to detect faces despite their size differences and low accuracy. The main contributions of this study are summarized as follows.

- Introduce a structurally modified YOLOv5 backbone that embeds attention directly within the feature extraction process. This is achieved by replace in original CSP1 module in the backbone with ISAM designed to better highlight the critical information, thereby facilitating the extraction of more relevant features. Additionally, every Convolutional BatchNormalize LeakyReLU (CBL) present in the backbone was replaced with a Convolutional BatchNormalize SiLU (CBS).
- In addition to the level of feature fusion layers that combine feature maps from different levels, these layers are augmented to collect more specific details about the small faces. Thus, four predictive heads are generated, which significantly reduce the impact of changes in object size and improve the ability to detect smaller objects in the Neck part.
- To improve the multi-perspective spatial contexts through parallel pooling operations. The ISAM operations module (median pooling, average pooling, and max pooling) was processed through a dedicated Convolution-BatchNorm-SiLU (CBS) block. This ensures that important information from different pooling methods is preserved and highlighted, resulting in a richer and more robust feature representation.

2. METHODS

In this section, we first provide a brief introduction to the deep learning-based detection method for masked face detection, as well as an introduction to the attention mechanism. Finally, we describe the CSPDarknet model and its improvements.

2.1 The YOLOv5 method

The YOLOv5 architecture consists of components: Backbone “CSPDarknet”, Neck, and Output, which are responsible for extracting informative features from the input images and combining these features to produce three sets of feature maps at different scales. Finally, the Output, which is the last part of the network, utilizes multiscale feature maps generated by the neck to predict the presence and location of objects in the input image. The model architecture includes a convolutional neural network as the backbone, designed from the input image, extract multi-scale feature maps through successive convolutional and pooling operations. The backbone produces four layers of feature maps with varying dimensions. These multi-scale maps are then processed by the neck network, which merges them to enhance contextual information and prevent data loss, using Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) structures. The FPN transfers strong semantic features from the higher to lower layers, whereas the PAN improves localization features by moving information from lower to higher layers. Together, these structures significantly enhance the feature fusion capability of the neck network. This fusion generated three scales of new feature maps: 76 × 76 × 255, 38 × 38 × 255, and 19 × 19 × 255, where 255 denotes the number of channels. These different sizes enable the detection of objects at various scales: larger objects from the 19 × 19 × 255 maps, and smaller objects from the 76 × 76 × 255 maps. The final detection and classification of objects occur in the output network.

2.2 The Spatial Attention Module (SAM)

The performance of deep learning models can be significantly improved by incorporating a spatial attention mechanism, particularly in image-processing and object-detection tasks. This mechanism operates by concentrating on specific regions of the input image that are deemed vital for the task at hand, thereby enabling the model to focus on the areas containing critical information [25]. For example, in object detection, a spatial attention mechanism can significantly improve the detection accuracy of smaller or less prominent objects by ensuring that these regions receive more attention during the feature-extraction process. Mathematically, two maps are created, each of which indicates the average pool features F_{avg} and maximum pool features F_{max} across the channel. they were combined using a convolutional layer to obtain a two-dimensional (2D) spatial attention map. Briefly, it is calculated as follows [19].

$$M_S = \sigma\left(f^{7\times7}\left(\left[F_{avg;}F_{max}\right]\right)\right) \tag{1}$$

where the equation symbols indicate the following: M_s to the result of the spatial attention mechanism, σ to the sigmoid function, 7 × 7 and to the filter size for the convolution operation f.

2.3 Improved CSPDarknet with spatial attention mechanism

To detect faces, we improved the detection method of CSPDarknet using an attention mechanism. As illustrated in Figure 1, which has several improvements to the YOLOv5 model and the SAM: (1) The CBL module was replaced by CBS in the architecture. (2) The improved SAM was integrated into the backbone instead of the CSP1_X module. (3) A new feature fusion layer was added and CSP2_X was replaced with a C3 × 3 unit in the neck. (4) The SAM is improved by separating and increasing the number of pooling channels.

First, compared with the original YOLOv5 architecture, the CBL in the backbone was replaced by CBS to enhance the feature representation ability. Note that CBL consists of "convolutional layer, batch normalization (BN), and leakyrelu activation function", whereas for CBS, it is only the activation function replaced by The Sigmoid Linear Unit, and thus it becomes as follows (Convolutional Layer, BN, The Sigmoid Linear Unit (SiLU) Activation Function). The main importance of this replacement is the improvement offered by SiLU activation, as it combines the properties of both linear and nonlinear activation, allowing it to smoothen the activation function more than LeakyReLU. This smoothness can contribute to improving the deep learning process of the CSPDarknet model by reducing the problem of vanishing gradients, which is a common problem in deep neural networks. In addition, it improves the prediction accuracy and reduces the error in the object detection process, especially in complex tasks, such as detecting occluded faces or small objects in images. Therefore, every CBL used in the architecture is replaced by a CBS, as in an SPP module and neck. As in Figure 1.

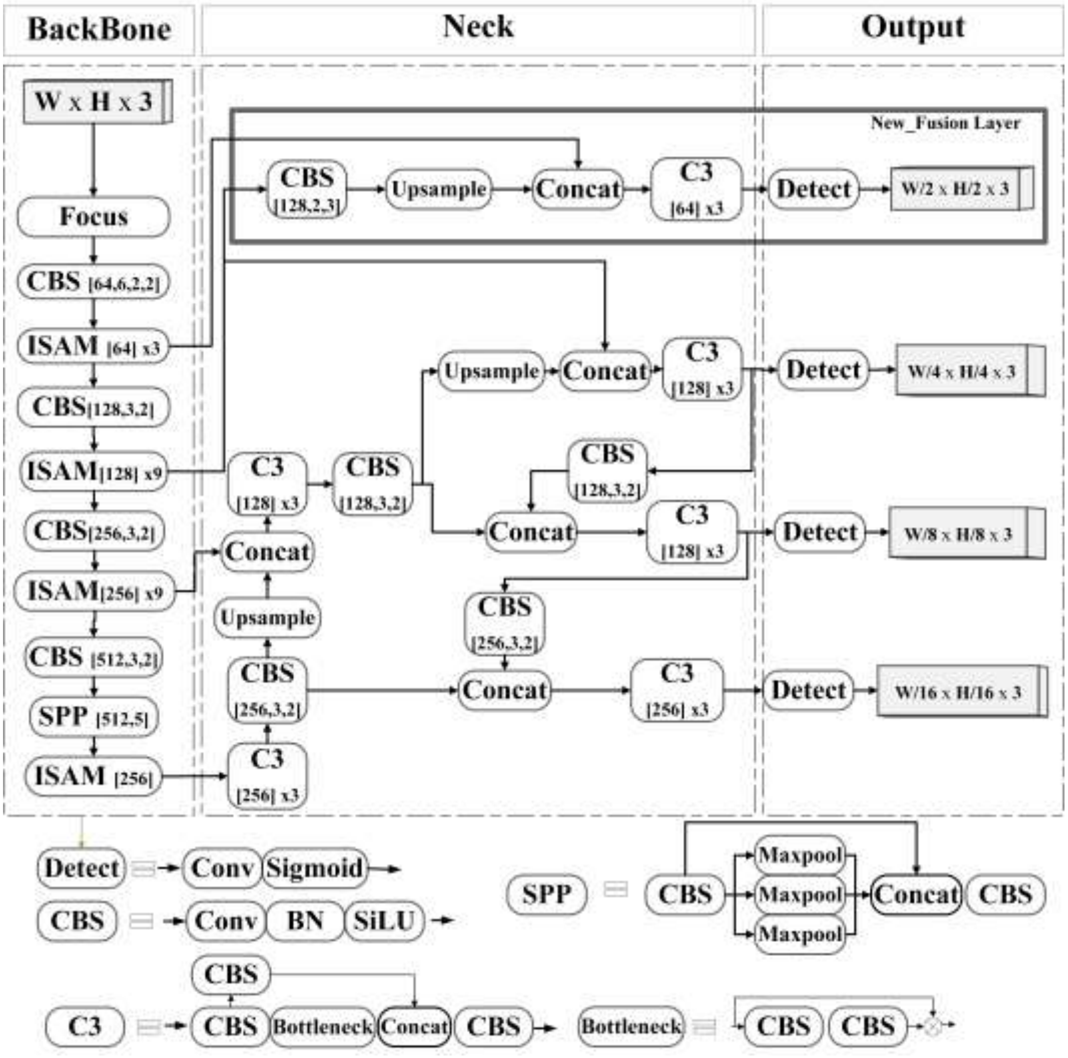

Figure 1. Proposed of YOLO-ISAM architecture

Second, in the backbone part, the CSP1_X module was replaced with an improved spatial attention mechanism (ISAM) to focus on important features and ignore weak features to obtain more accurate and faster results. Integrating the spatial attention mechanism into this site instead of the CSP1_X unit in the backbone represents a strategic improvement that enhances the accuracy and effectiveness of the model. The process of extracting features starts from very early stages, and because of its ability to focus on important areas and ignore unnecessary information, it contributes to the mechanism of spatial attention in reducing the loss of vital information during the pooling and miniaturization that occur in multiple stages of the network.

Third, for the neck part, this includes adding a new feature fusion layer and links or paths to the structure to create a map of features larger than 152 × 152 × 255. The fusion layer operations include unsampling and concatenation, which are fed from the backbone after adding the ISAM. and CSP2_X was replaced with a C3 × 3 unit. The C3*3 module (referring to the use of a triple convolutional layer sequence of 3×3 size). Using a sequence of 3 × 3 convolutional layers allows the model to better handle spatial variations in the image, and the model may obtain better and more stable gradients during the training process. Although C3*3 may increase the number of calculations compared to the CSP2_X unit, it achieves a better balance between accuracy and efficiency by improving the exploitation of gradients and information flow over the network. In addition, replacing the CSP2_X module with the C3 × 3 module improved the inference speed by reducing the size of the model without sacrificing its ability to identify valuable visual features. To clarify, CSP2_X differs from CSP1_X in its backbone in that it replaces the remaining network with a 2 × X CBS to enhance its ability to integrate network features. where X is the number of units. Figure 2 shows the improved CSPDarknet that apply on Yolov5 architecture.

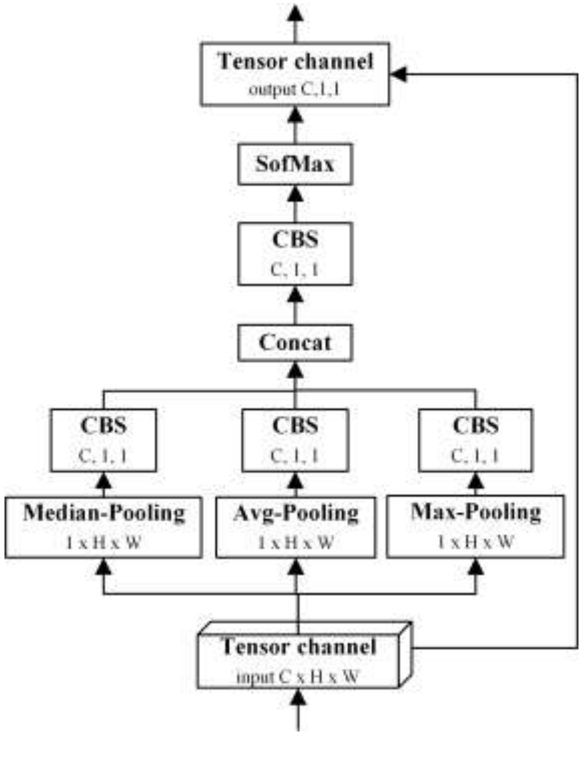

Figure 2. Improve Spatial attention Module (ISAM) structure

Fourth, to focus on the small amount of information visible from the face, an improved attention mechanism was added to the backbone of the YOLOv5 model, as illustrated in Figure 2. The importance of incorporating the attention mechanism module at this location into the backbone replacement with the CSP1_X module represents a strategic improvement that contributes to enhancing the accuracy and effectiveness of the model. In terms of the feature extraction process that starts at very early stages and owing to its ability to focus on important regions and eliminate unnecessary information, the spatial attention mechanism contributes to reducing the loss of vital information that occurs at multiple stages of the network. The SAM adopts three separate operations to generate three maps: average, maximum, and average pooling. This is followed by the CBS for each map to effectively utilize contextual information, and then combined with CBS.

Finally, the sigmoid function is used to obtain the final spatial attention map A. The use of multiple pooling channels, such as median pooling, average pooling, and max pooling, allows the model to accommodate a variety of spatial features from the image. Each type of pooling captures different aspects of the data, enhancing the model’s understanding of the fine and varied details in the image. In particular, using median pooling helps reduce noise from outliers in the data, resulting in a higher stability in the model’s response. While Avg-Pooling provides an average estimate of the features at the pixel level, max-pooling captures the sharpest and most robust features. In addition, separating the pooling channels allows each to focus on specific aspects of the image without being affected by the others.

The block diagram of the SAM is illustrated in Figure 2 The spatial attention mechanism can be mathematically represented as follows the algorithm 1 for ISAM.

Algorithm 1: ISAM

Input: Tensor channel input X

Output: Tensor Attention-weighted output Y

```
Begin:
Step 1: Perform Median-Pooling, Avg-Pooling, and Max-Pooling on input tensor X:
    Xmed=medianPooling(X)
    Xavg=avgPooling(X)
    Xmax=maxPooling(X)
Step 2: Pass the pooled tensors through their respective Convolution-Batch Normalization- SiLU (CBS) blocks:
    CBSmed=CBS(Xmed)
    CBSavg=CBS(Xavg)
    CBSmax=CBS(Xmax)
Step 3: Concatenate the outputs of the CBS blocks:
ConcatCBS=Concat(CBSmed,CBSavg,CBSmax)
Step 4: Pass the concatenated result through another CBS block:
    CombinedCBS=CBS(ConcatCBS)
Step 5: Apply SoftMax function to the output of the final CBS block to generate the attention map:
    A=SoftMax(CombinedCBS)
Step 6: Multiply the original input tensor X element-wise with the attention map A to obtain the final output tensor Y:
    Y= A × X
Step 7: Return the final output tensor Y
End
```

2.4 Detailed implementation of the ISAM module

To enhance the feature representation capabilities of the backbone network, we introduce the Improved ISAM. The ISAM leverages multiple pooling strategies to generate a robust attention map. Figure 2. illustrates the overall structure, and the detailed forward pass is formulated as follows. Given an input feature tensor X ∈ $R^{C\times H\times W}$, the ISAM module computes the output Y through the following steps:

- Multi-Branch Pooling is a three pooling operations in parallel to capture different spatial contexts:

Median Pooling $X_{med} = MedianPool(X),$ with a kernel size of K ×K and a stride of s. this operation highlight robust.

Average Pooling $X_{avg} = AvgPool(X),$ with a kernel size of K ×K and a stride of s. this operation captures the mean spatial context.

MaxPooling $X_{max} = MaxPool(X),$ with a kernel size of K × K and a stride of s. this operation emphasizes the salient features.

Parameter kernel size of k=3 and stride s=1 was chosen for all pooling operations. This size is enough to capture local spatial relationships without causing excessive loss of resolution. and also use 'same' padding to maintain the spatial dimensions H ×W.

- Feature Transformation each pooled feature map is then passed through an independent (CBS) block.

$$CBS_{med} = CBS(X_{med})$$
$$CBS_{avg} = CBS(X_{avg})$$
$$CBS_{max} = CBS(X_{max})$$

Each convolution in these blocks has kernel size of 3 ×3, a stride of 1, and ‘same’ padding. The number of output channels for each convolution is set to C, which is typically C/4 or C/2 to create a bottleneck and reduce computational overhead. The batch-normalization layer uses a momentum 0.03 and an epsilon of 1e-3.

- Feature fusion and attention map generation are concatenated the three output CBS blocks channel dimension.

$$Concat_{CBS} = [CBS_{med}, CBS_{avg}, CBS_{max}] \in R^{3C\times H\times W}$$

Thus fused tensor is then passesd through a final CBS block with 1 × 1 convolution to compress the channels back to C.

$$Combined_{CBS} = CBS_{1\times1}(Concat_{CBS}) \in R^{C\times H\times W}$$

Thus is A SoftMax function is applied spatially to normalize the values into an attention map A where each location (i, j) sums to 1.

$$A_{c,i,j} = \frac{exp\ (Combined_{CBS_{c,i,j}})}{\sum_{h=1}^{W}\sum_{w=1}^{W} exp\ (\ Combined_{CBS_{c,h,w}})} \quad (2)$$

$$Y = A \cdot X \quad (3)$$

Feature Recalibration is final output Y is obtained by performing an element-wise multiplication between the original input X and the attention map A, effectively re-weighting the importance of each spatial location, computational Complexity in ISAM module is operations are the convolutions within the CBS blocks.

3. EXPERIMENTAL SETUP

3.1 Dataset

In this study, four sets of video datasets were used: ChokePoint [26] and ViDMASK [27]. Studying a video database is important to produce effective mechanisms for detecting masked faces and recognizing an individual's identity. For the third dataset, Moxa3K [28], which we used as a standard for detecting faces because it is diverse and contains unclear and crowded samples and different lighting conditions, and the LFW-SM [29] dataset for mask face facial image. Table 1. shows the sources of the database.

Table 1. Lists the sources of the databases

No.	Dataset	Number of Images / Video
1	ChokePoint [26]	48 Video
2	VIDMASK [27]	67 Videos
3	Moxa3k [28]	3000 Image
4	LFW-SM [29]	13233 Image

Table 2. Details of the dataset used

Dataset	Total- image	Train	Valid	Test
Chockpoint	3166	2294	290	582
VIDMASK	7295	5120	727	1448
Moxa3k	2928	2049	293	586
LFW-SM	13233	9263	1324	2646

The Chokepoint Dataset is a real-world surveillance dataset designed for person identification and verification experiments. It includes videos captured using three cameras placed above choke points for pedestrian traffic, resulting in a dataset that simulates real-world scenarios. The dataset includes face images captured while a person is walking through a portal, with variations in illumination conditions, pose, sharpness, and misalignment owing to automatic face localization and detection. The Chokepoint Dataset consists of 25 individuals (19 males and 6 females) in Portal 1 and 29 individuals (23 males and 6 females) in Portal 2, totaling 48 video sequences and 64,204 face images. Each sequence is named based on the recording conditions, such as the portal, sequence, and camera labels.

The VIDMASK dataset includes videos of unidentified individuals wearing or not wearing masks in various crowds and incidental scenarios. Out of 67 videos. The frames were annotated and shuffled. A total of 20,000 instances of masks and 2,500 non-mask-wearing individuals were identified in the images. The VIDMASK dataset was obtained from “YouTube” and “Pexel.com”. The videos displayed real environments, with being interviewed and working on daily tasks, resulting in real poses expressions.

The Moxa3K dataset consists of approximately 3000 images. It contains images of many people and close-up shots of the people's profiles. Additionally, it includes Google search images of the population during the pandemic. What will be mentioned now applies to all the datasets used. The datasets were divided into training, testing, and validation sets at ratios of 70, 20, and 10%, respectively.

The Labelled Faces in the Wild simulated masks (LFW-SM) dataset is contains images with the simulated mask applied as a standard benchmark dataset used to evaluate the performance of face recognition systems. It contains 5749 identities with a total of 13233 images. As mentioned earlier, the objective of masked face recognition is to accurately identify individuals both with and without masks [30, 31]. The images in the test pairs are selected across the datasets to evaluate the robustness of the network. Partitioning was performed as shown in Table 2 for all datasets.

3.2 Evaluation metrics

To determine the mAP, precision, and recall, the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP refers to the images that were correctly classified as true, whereas TN denotes the images that were correctly classified as false. On the other hand, FP refers to results that have been incorrectly classified as false but predicted correctly, whereas FN refers to images that have been correctly classified as true but predicted as false. In the context of mask detection, TP, FN, FP, and TN were measured based on the objectless of the detection result, as determined through the intersection of the union measurement. Intersection over Union (IoU) is a widely used metric for evaluating the accuracy of regression in object detection models. It measures the overlap between a predicted bounding box and its corresponding ground-truth box [28]. An IoU equal to or greater than 50% is classified as a “True Positive (TP)”, indicating a correct detection, whereas an IoU below 50% is classified as a “False Positive (FP)”, indicating an incorrect detection. “False Negatives (FN)” were determined by counting objects that were not detected. To measure the percentage of correct predictions, precision and recall were calculated using Eqs. (4)-(7).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$AP = \int_0^1 P(R)\, dR \tag{6}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{7}$$

P, R, and N represent the precision, recall rate, and count of all the objects in each category, respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

3.3 Validation strategy

We used 5-fold cross-validation to rigorously evaluate our model and prevent overfitting. The data was first split into five separate groups of equal size, making sure the proportion of each class was maintained in every group. We then trained and tested the model five separate times. For each run, we used four groups for training and held out the fifth for testing. This process guaranteed that every data point was used for testing once and only once. Our final results are the average and variation of these five runs, which gives us a much more trustworthy measure of how the model will perform on new data compared to a simple single split.

4. COMPARISON AND ANALYSIS OF EXPERIMENT

This proposed conducted a comparative performance analysis between the proposed YOLO-ISAM framework and existing YOLO series models across three benchmark datasets: ChokePoint, VIDMASK, Moxa3K, and LFW-SM. As summarized in Table 3, YOLO-ISAM demonstrated superior detection capabilities, particularly on the ChokePoint dataset, where it achieved a substantial improvement in mean AP at 50% intersection-over-union (mAP@50) compared to baseline models.

The experimental results validate the effectiveness of the proposed architectural enhancements in addressing the challenges of masked face detection.

Table 3. Comparison of the performance of the YOLO series with our model

Dataset	Chokepoint			VIDMASK			Moxa3k			LFW-SM		
Methods	mAP@0.5	Accuracy	F1-Score	mAP@0.5	Accuracy	F1-Score	mAP@ 0.5	Accuracy	F1-Score	mAP@ 0.5	Accuracy	F1-Score
YOLOv3	0.35	0.40	0.38	0.50	0.58	0.55	0.63 [32]	0.66	0.65	0.94	0.96	0.95
YOLOv4	0.36	0.40	0.39	0.49 [33]	0.55	0.53	0.68 [34]	0.70	0.69	0.94	0.95	0.95
YOLOv5	0.45	0.48	0.48	0.86	0.90	0.88	0.65 [32]	0.66	0.65	0.98	0.98	0.98
YOLOv6	0.44	0.46	0.45	0.81	0.82	0.81	0.64	0.67	0.67	0.97	0.98	0.97
YOLOv7	0.48	0.52	0.50	0.77	0.81	0.89	0.66	0.68	0.68	0.96	0.99	0.98
YOLOv8	0.56	0.63	0.61	0.88	0.90	0.92	0.69	0.75	0.74	0.94	0.97	0.95
YOLOv9	0.66	0.72	0.73	0.90	0.93	0.91	0.68	0.74	0.73	0.98	0.99	0.99
YOLOv10	0.58	0.65	0.62	0.93	0.95	0.94	0.68	0.74	0.75	0.99	0.99	0.98
YOLOv11	0.78	0.81	0.78	0.94	0.96	0.96	0.79	0.88	0.86	0.96	0.98	0.97
YOLOv12	0.79	0.81	0.79	0.91	0.93	0.91	0.78	0.83	0.80	0.98	0.99	0.99
YOLO-ISAM	0.86	0.92	0.91	0.96	0.99	0.99	0.84	0.91	0.90	0.99	1.00	0.99

Table 4. A comparison of our proposed network with different ablation results of baseline YOLOV5 And Improve YOLOV5

Dataset	Chokepoint			VIDMASK			Moxa3K			LFW-SM		
Methods	mAP@0.5	Accuracy	F1-Score	mAP@0.5	Accuracy	F1-Score	mAP@0.5	Accuracy	F1-Score	mAP@0.5	Accuracy	F1-Score
VGG-16	0.32	0.35	0.34	0.79	0.80	0.79	0.46	0.48	0.45	0.97	0.98	0.97
ResNet-18	0.44	0.45	0.45	0.85	0.85	0.84	0.62	0.63	0.63	0.98	0.99	0.99
Baseline (CSPDarknet53)	0.38	0.37	0.34	0.81	0.85	0.81	0.62	0.58	0.53	0.96	0.97	0.95
CSPDarknet53 With ISAM	0.40	0.40	0.38	0.83	0.86	0.85	0.63	0.60	0.57	0.99	0.99	0.99
Improve- YOLOv5s with 4 -Fusion Layer	0.42	0.47	0.46	0.87	0.91	0.90	0.66	0.61	0.61	0.98	0.98	0.97

5. ABLATION STUDY

An ablation study was conducted to systematically evaluate the contribution of each component in the proposed YOLO-ISAM architecture. As detailed in Table 4, we incrementally integrated key modifications—including the ISAM and additional fusion layers—into the baseline CSPDarknet53 backbone. The experimental framework involved selectively removing architectural elements to isolate their individual impact on detection performance across all three datasets. Furthermore, statistical significance testing using Student's t-test was performed to validate the observed performance improvements, with results presented in Table 5. The consistent improvement observed across all cross-validation folds, supported by preliminary t-test results, suggests a strong performance advantage for our model. Future work with more extensive testing will be needed to confirm statistical significance with high power.

Table 5. Comparison of YOLOv12 models with YOLO-ISAM models as a Statistical calculation an independent samples t-test

Dataset	Round	YOLOv12		YOLO-ISAM		T-Value	P-Value
		Mean %	STD %	Mean %	STD %		
Chokepoint	25	80.85	1.64	88.42	3.73	9.28	0.0033
VIDMASK	25	93.4	1.71	97.42	2.33	6.95	0.0012
Moxa3K	25	84.26	1.05	88.72	2.40	8.5	0.0018
LFW-SM	25	95.4	2.0	99.4	3.3	4.69	0.00481

6. CONCLUSIONS

This paper introduces YOLO-ISAM, an improved method to detect the face mask recognition. The proposed model was integrates on improved Spatial Attention Mechanism (ISAM) directly into the YOLOv5 backbone (CSPDarknet-53), a structural modification that compels the network to learn feature representations prioritizing the most salient visible facial regions. Furthermore, enhancements to the new feature fusion network improved multi-scale detection capabilities, which are critical for identifying faces at a distance. Comprehensive on four databases, two of which are video and the last are images, which were used as standards, Chockpoint, VIDMASK, Moxa3k, and LFW-SM, respectively. Experimental results show that compared with VGG-16, Resnet-18, YOLO series, and YOLO-ISAM. In addition, there are still significant challenges in recognizing masked faces, such as the lack of datasets and the complexity of face occlusion. Furthermore, the model's performance is inherently tied to the diversity of the available training data; a lack of datasets representing an exhaustive range of mask types, ethnicities, and extreme lighting conditions remains a challenge for the field. In future work, we plan to create a special dataset for masked face recognition. Simultaneously, we can learn identity from multiple aspects, including voice, to improve recognition accuracy.

REFERENCES

[1] Hsu, G.S.J., Wu, H.Y., Tsai, C.H., Yanushkevich, S., Gavrilova, M. L. (2022). Masked face recognition from synthesis to reality. IEEE Access, 10: 37938-37952. https://doi.org/10.1109/ACCESS.2022.3160828

[2] Himeur, Y., Al-Maadeed, S., Varlamis, I., Al-Maadeed, N., Abualsaud, K., Mohamed, A. (2023). Face mask detection in smart cities using deep and transfer learning: Lessons learned from the COVID-19 pandemic. Systems, 11(2): 107. https://doi.org/10.3390/systems11020107

[3] 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

[4] Saravanan, T.M., Karthiha, K., Kavinkumar, R., Gokul, S., Mishra, J.P. (2022). A novel machine learning scheme for face mask detection using pretrained convolutional neural network. Materials Today: Proceedings, 58: 150-156. https://doi.org/10.1016/j.matpr.2022.01.165

[5] Carragher, D.J., Towler, A., Mileva, V.R., White, D., Hancock, P.J.B. (2022). Masked face identification is improved by diagnostic feature training. Cognitive Research: Principles and Implications, 7: 30. https://doi.org/10.1186/s41235-022-00381-x

[6] Wu, P., Li, H., Zeng, N., Li, F. (2022). FMD-Yolo: An efficient face mask detection method for COVID-19 prevention and control in public. Image and Vision Computing, 117: 104341. https://doi.org/10.1016/j.imavis.2021.104341

[7] Cabani, A., Hammoudi, K., Benhabiles, H., Melkemi, M. (2021). MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19. Smart Health, 19: 100144. https://doi.org/10.1016/j.smhl.2020.100144

[8] Santosh, K.C., Das, N., Ghosh, S. (2022). Deep learning: A review. In Deep Learning Models for Medical Imaging, pp. 29-63. https://doi.org/10.1016/B978-0-12-823504-1.00012-X

[9] Pan, Y., Zhang, G., Zhang, L. (2020). A spatial-channel hierarchical deep learning network for pixel-level automated crack detection. Automation in Construction, 119: 103357. https://doi.org/10.1016/j.autcon.2020.103357

[10] Huang, D.Y., Chen, C.H., Chen, T.Y., Hu, W.C., Guo, Z.B., Wen, C.K. (2020). High-efficiency face detection and tracking method for numerous pedestrians through face candidate generation. Multimedia Tools and Applications, 80(1): 1247-1272. https://doi.org/10.1007/s11042-020-09780-y

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

[12] Nazir, A., Wani, M.A. (2023). You only look once-object detection models: a review. In 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 1088-1095.

[13] Redmon, J., Farhadi, A. (2017). YOLO9000: Better,

faster, stronger. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, pp. 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[14] Redmon, J., Farhadi, A. (2018). YOLOV3: An incremental improvement. arXiv preprint arXiv:1804.02767. https://doi.org/10.48550/ARXIV.1804.02767

[15] Kumar, S., Vishal, Sharma, P., Pal, N. (2021). Object tracking and counting in a zone using YOLOv4, DeepSORT and TensorFlow. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, pp. 1017-1022. https://doi.org/10.1109/ICAIS50930.2021.9395971

[16] Golwalkar, R., Mehendale, N. (2022). Masked-face recognition using deep metric learning and FaceMaskNet-21. Applied Intelligence, 52(11): 13268-13279. https://doi.org/10.1007/s10489-021-03150-3

[17] Nepal, U., Eslamiat, H. (2022). Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. Sensors, 22(2): 464. https://doi.org/10.3390/s22020464

[18] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[19] Wang, Y., Li, Y., Zou, H. (2023). Masked face recognition system based on attention mechanism. Information, 14(2): 87. https://doi.org/10.3390/info14020087

[20] Woo, S., Park, J., Lee, J.Y., Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Computer Vision - ECCV 2018: 15th European Conference, Munich, Germany, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[21] Wang, L., Cao, Y., Wang, S., Song, X., Zhang, S., Zhang, J., Niu, J. (2022). Investigation into recognition algorithm of helmet violation based on YOLOv5-CBAM-DCN. IEEE Access, 10: 60622-60632. https://doi.org/10.1109/ACCESS.2022.3180796

[22] Rajamohanan, R., Latha, B. C. (2023). An optimized YOLO v5 model for tomato leaf disease classification with field dataset. Engineering, Technology & Applied Science Research, 13(6): 12033-12038. https://doi.org/10.48084/etasr.6377

[23] Zhang, X., Shang, S., Tang, X., Feng, J., Jiao, L. (2022). Spectral partitioning residual network with spatial attention mechanism for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-14. https://doi.org/10.1109/TGRS.2021.3074196

[24] Montero, D., Nieto, M., Leskovsky, P., Aginako, N. (2022). Boosting masked face recognition with multi-task ArcFace. In 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Dijon, France, pp. 184-189. https://doi.org/10.1109/SITIS57111.2022.00042

[25] Ge, Y., Liu, H., Du, J., Li, Z., Wei, Y. (2023). Masked face recognition with convolutional visual self-attention network. Neurocomputing, 518: 496-506. https://doi.org/10.1016/j.neucom.2022.10.025

[26] ChokePoint Dataset. https://zenodo.org/records/815657, accessed on Sep. 15, 2025.

[27] GitHub - ViDMask/VidMask-code. https://github.com/ViDMask/VidMask-code?tab=readme-ov-file, accessed on Aug. 15, 2025

[28] MOXA: A Deep Learning Based Unmanned Approach For Real-Time Monitoring of People Wearing Medical Masks. https://shitty-bots-inc.github.io/MOXA/index.html.

[29] Anwar, A., Raychowdhury, A. (2020). Masked face recognition for secure authentication. arXiv preprint arXiv:2008.11104. https://doi.org/10.48550/arXiv.2008.11104

[30] Ding, F., Peng, P., Huang, Y., Geng, M., Tian, Y. (2020). Masked face recognition with latent part detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, USA, pp. 2281-2289. https://doi.org/10.1145/3394171.3413731

[31] Mahmoud, M., Kasem, M. S., Kang, H.S. (2024). A comprehensive survey of masked faces: Recognition, detection, and unmasking. Applied Sciences, 14(19): 8781. https://doi.org/10.3390/app14198781

[32] Oday, A., Abdullah, A., Sahran, S. (2025). YOLO-OSAM: Reassembly spatial attention mechanisms for facial expression recognition. Traitement du Signal, 42(4): 2379-2387. https://doi.org/10.18280/ts.420445

[33] Ottakath, N., Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Mohamed, A., Khattab, T., Abualsaud, K. (2022). ViDMASK dataset for face mask detection with social distance measurement. Displays, 73: 102235. https://doi.org/10.1016/j.displa.2022.102235

[34] Roy, B., Nandy, S., Ghosh, D., Dutta, D., Biswas, P., Das, T. (2020). MOXA: A deep learning based unmanned approach for real-time monitoring of people wearing medical masks. Transactions of the Indian National Academy of Engineering, 5(3): 509-518. https://doi.org/10.1007/s41403-020-00157-z