Facial Emotion Recognition in Occluded Images Using Attention-Based CNN

Anil Rana[1], Ashim Khadka[2]*, Jianchu Lin[3], Lin Zhang[4]

[1]Department of Graduate Studies, Gandaki College of Engineering and Science, Pokhara, Nepal;
msc2019ise27@gces.edu.np

[2]*Faculty of Science and Technology, Nepal College of Information Technology, Lalitpur, Nepal;
ashim.khadka@ncit.edu.np

[3]Faculty of Computer and Software Engineer, Huaiyin Institute of Technology,
Jiangsu, China; linjianchu@hyit.edu.cn

[4]Faculty of Management Engineering, Huaiyin Institute of Technology, Huaian, China; zlmjl@hyit.edu.cn

*Corresponding Author: ashim.khadka@ncit.edu.np

DOI: https://doi.org/10.30212/JITI.202503.008

Submitted: Dec. 30, 2024 Accepted: Apr. 01, 2025

ABSTRACT

Facial expression recognition (FER) is widely used in various applications, yet few studies address its effectiveness under occlusion conditions. Occlusions can obscure critical facial features, leading to the loss of valuable expression information and negatively impacting recognition performance. This study enhances the robustness of FER models by integrating both channel and spatial attention mechanisms into a convolutional neural network (CNN). The attention module improves feature extraction by selectively focusing on visible facial regions while compensating for missing information, thereby enhancing recognition performance in obscured facial images. The proposed model is evaluated on both synthetic and real-world occlusion datasets, including RAF, FED-RO, CK+, JAFFE, FER2013, and AffectNet, demonstrating its robustness across different occlusion scenarios. Experimental results show that the prposed model achieves an accuracy rate of 66%, outperforming several state-of-the-art methods. Additionally, cross-dataset and k-fold validation confirm the model's generalization capabilities across different datasets and occlusion patterns, further validating its reliability in real-world applications. The results demonstrate that attention-based CNNs effectively mitigate occlusion effects and improve emotion classification.\.

Keywords: Facial expression recognition, Occlusion, CNN, Channel-Spatial attention

1. Introduction

Human communication has been a significant research topic for the past decades. With advancements in technology, various approach has been proposed over time to recognize human emotion in different way. Facial expressions contain rich personal emotional information and the automatic recognition of expressions has broad application prospects in the fields of human-computer interactions, intelligent security, psychological analysis, behaviour prediction and interpersonal relations[1], [2], [3], [4], [5], [6]. Although several studies have been conducted on facial expression recognition (FER), identifying expressions independent of pose, face shape, image resolution, brightness, and occlusion remains a challenging task [7], [8], [9], [10].

Occlusion is one of the key challenges that limit the information available in an image, significantly compromise the recognition ability of the model. The facial expression images captured by image acquisition device often have partial occlusion which commonly caused by hands, glasses, masks, etc as shown in fig.Fig *1*. Real examples with different types of facial occlusion [12] Occlusions can obstruct important facial features such as the eyes and mouth, interfering with emotion recognition and affecting expression recognition accuracy. Facial occlusion is considered one of the most challenging problems that degrade the performance of the face recognition system such as surveillance, healthcare, and human-computer interaction [11], [12], [13]. Therefore, developing a model that can effectively recognize facial expressions under occlusion conditions is crucial for improving performance in practical applications.

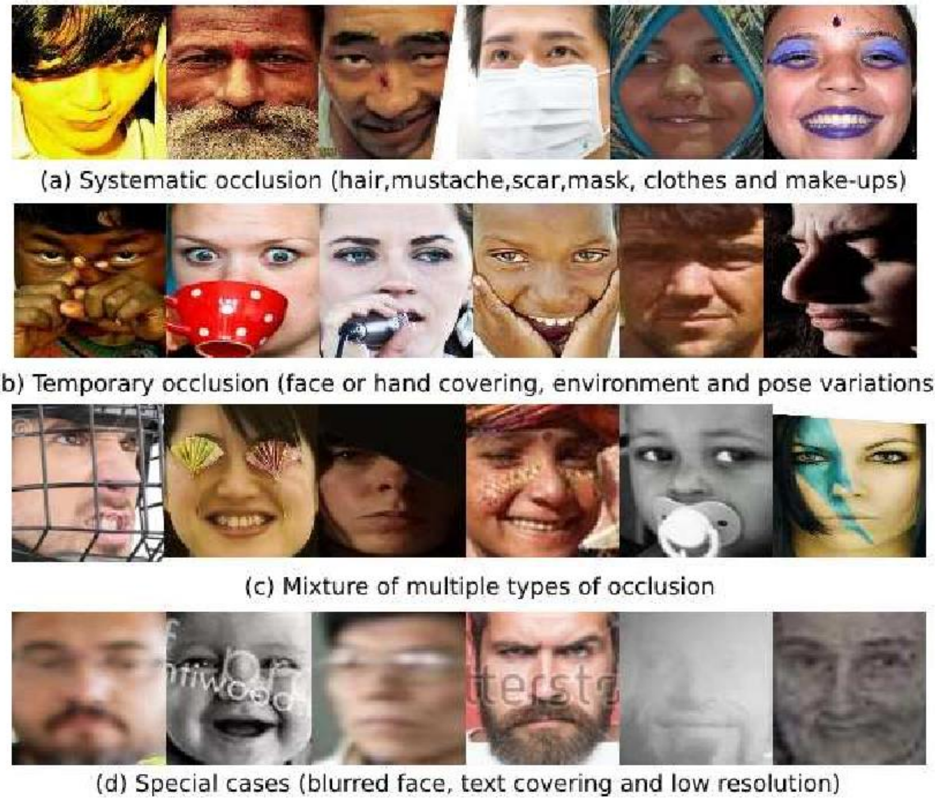
(a) Systematic occlusion (hair,mustache,scar,mask, clothes and make-ups)
(b) Temporary occlusion (face or hand covering, environment and pose variations)
(c) Mixture of multiple types of occlusion
(d) Special cases (blurred face, text covering and low resolution)

Fig 1. Real examples with different types of facial occlusion [12]

Convolutional neural networks (CNNs) are regarded as robust automatic feature extractors that achieve good results in image recognition tasks [1], [2], [3], [14]. However, CNNs struggle in the presence of occlusions because CNNs extract features from the entire image without distinguishing between occluded and unoccluded regions. In contrast, the human visual system (HVS) does not process a whole image at once but focuses on specific parts to better capture visual structures. Attention mechanisms in neural networks focus on the important features while ignoring irrelevant ones. Attention mechanism has been proven useful in pixel-wise computer vision tasks and are used to measure how much attention to pay to the features in different regions [15], [16]. Attention-based feature refinement with two distinctive modules channel and spatial can achieve considerable performance improvements in classification and detection performances with various models [11], [17], [18], [19]. Recent work in occlusion-aware FER has explored Patch-Gated CNNs and similar attention-based mechanisms [12], [20], [21], [22]. Patch-Gated CNNs focus on localized patches of facial expressions to address occlusions; however, face limitation when occlusions are spread across multiple regions of the face. While these models show potential in identifying occluded areas, they struggle to effectively capture both the "what" (relevant features) and "where" (spatial locations) aspects of important features across the entire image.

This paper proposes a CNN model with an integrated attention module, utilizing both channel attention and spatial attention, as illustrated in Fig Figure *2*. CNN model with attention module, enabling it to selectively emphasize non-occluded regions and improve robustness to missing facial features. The Channel Attention Module enhances feature selection by exploiting inter-channel relationships of features, effectively determining ‘what’ is meaningful in an input image. To compute channel attention efficiently, the spatial dimensions of the input feature map are squeezed. The Spatial Attention Module focuses on the inter-spatial relationships of features, identifying ‘where’ the most informative regions are located within an image. Since important features can appear in different positions across multiple instances, spatial attention complements channel attention [18], [19]. By integrating both attention mechanisms, the model selectively attend to specific features, i.e., unoccluded regions which are important for the expression recognition of a facial image, leading to better performance. The two module emphasize on meaningful features along with channel and spatial dimensions, ensuring that each branch learns both ‘what’ and ‘where’ to focus on. This results in efficient information flow within the network, helping the model suppress irrelevant information and improve classification robustness. This hybrid attention mechanism allows the network to adapt dynamically to occlusion patterns, enhancing performace of recognition.

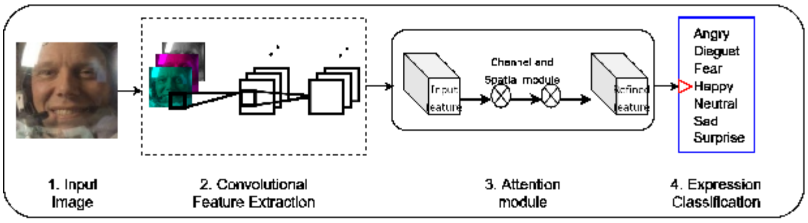

Figure 2. CNN model with attention module

The major contributions of this paper can be summarized as follows:

a) The use of the attention module to improve CNN performance in recognizing facial expressions in obscured facial images. This approach enhances feature extraction and focuses on occluded facial emotion aspects by incorporating both a channel attention module and a spatial attention module.

b) The model is evaluated on both real and synthetically occluded images from well-known datasets: RAF, FER-RO, CK+, JAFFE, FER2013, and AffectNet, demonstrating its robustness in handling occlusions.

c) The proposed model employes cross-dataset and k-fold validation to enhance its generalization capabilities across different datasets and occlusion patterns.

The rest of the paper is organized as follows: Section 2 discusses related work on occlusion-aware FER. Section 3 details the proposed methodology, while Section 4 discusses the detailed results. Finally, Section 5 discusses conclusion and future directions.

2. Literature Review

The facial expression recognition technique is capable of recognizing facial emotion with a high degree of accuracy. Recently, extensive efforts have been paid into the research towards overcoming the occlusion problem in face recognition systems. CNN methods are preferred for facial expression recognition as it has achieved great success in the past few years.

The CNN has combined Bidirectional Long Short-Term Memory Networks (BiLSTM) to leverage both spatial and temporal features of facial expressions, which is essential for accurate emotion recognition. Multi-task Cascaded Convolutional Networks (MTCNN) detect the face ensuring uniform dimension where BiLSTM is used to capture temporal dynamics of facial expressions. This dual approach enhances the model's ability to recognize emotions by addressing challenges such as variations in lighting, position, and expression intensity .The discriminative features are extracted from critical facial regions using a region-aware sub-net (RASnet) with multiple attention (MA). The RASnet uses binary masks to learn expression-related critical regions with coarse-to-fine granularity levels. The expression RASnet (eRASnet) with a multiple attention (MA) block learns comprehensive discriminative features. The proposed method recognizes the facial expression under non-occlusion conditions using publicly available databases, RAF and CK+ [23]. The down-sampling can lead to the loss of critical features that are essential for accurate emotion recognition. The integration of Squeeze-and-Excitation Networks (SENet) modules is proposed to maintain feature integrity, allowing for better feature extraction. By utilizing features from different layers of the network, the proposed method demonstrates the effectiveness of multi-task learning in deep learning frameworks. The experimental results showing an accuracy of 68.87% underscore the effectiveness of the proposed method [7].

In real-world situations when faces may be partially occluded by objects or accessories, the occlusions can significantly impair the model's capacity to accurately recognize expressions. The proposed attention-based ResNet-18 network shows increased accuracy and robustness in facial expression recognition, suggesting that data augmentation and attention mechanisms together can significantly enhance the performance of facial expression recognition systems in real-world applications [24]. Patch-gated CNN for facial expression recognition under occlusion consists of region decomposition, mimicking how humans recognize facial expressions. It automatically ignores the blocked facial patch and pays attention mainly to the unblocked and informative patches [20]. Wasserstein generative adversarial network-based method has also been used to perform occluded facial expression recognition. It consists of a generator G and two discriminators D1 and D2. The occluded part is reconstructed to form a complete non-occluded image that increases the FER system accuracy [25].

A deep residual network with CBAM is proposed to enhance feature extraction from partially occluded facial expression data. Multi-task Cascaded Convolutional Networks (MTCNN) is used to precisely localize key facial regions that convey emotional information. The CBAM-ResNet network extracts deep emotional features from the localized areas [26]. An occluded facial expression recognition method is proposed to utilise non-occluded images as privileged information to enhance the occluded classifier. Specifically, two deep neural networks are first trained from occluded and non-occluded facial images respectively. Then the non-occluded network is fixed and used to guide the fine-tuning of the occluded network from both label and feature space [27]. The occlusion-aware multi-scale attention consistency network (OMAC-Net) is designed to effectively capture both global and local fine-grained representations of facial expressions. Low-level feature random destruction (LFRD) module extracts rich local fine-grained information which helps to retain important edge information and the correlation between different patches of the face. Multi-scale attention consistency constraint (MACC) allows it to focus on the most discriminative features relevant for recognizing facial expressions [11].

The occlusion detection module based on symmetric SURF is presented to detect the occlusion part of the face. Then, the face inpainting module based on mirror transition is proposed to reconstruct the occluded area. Finally, a facial expression recognition network based on heterogeneous soft partitions is proposed to dynamically assign weights according to the proportion of the occluded part of each block, and the weighted facial images are fed into the trained neural network model for facial expression recognition.[28] Occlusion Adaptive Deep Network which consists of two branches: a Landmark-guided Attention Branch and a Facial Region Branch. A landmark-guided attention branch is used to find and discard corrupted features from occluded regions so that they are not used for recognition. An attention map is first generated to indicate if a specific facial part is occluded and guide the model to attend to non-occluded regions [29].

3. Research Methodology

3.1 Data Collection

The dataset consists in-the-lab (synthetics images) and in-the-wild generated (real images) which are publicly available. Affectnet [30], FER2013 [31], RAF-DB [12], FED-RO[32] datasets are collection of images captured in natural environment whereas JAFFE [33] and CK+ [34] datasets are lab-generated images. Facial images are classified into seven categories as shown in Table 1.

Distribution of data in seven classes.

Table 1. Distribution of data in seven classes

Emotion	FER2013	JAFFE	RAF-DB	CK+	FED-RO	Affectnet
Angry	4952	30	867	45	53	5000
Disgust	546	29	877	59	51	3803
Fear	5120	32	355	25	58	5000
Happy	8988	31	5957	69	59	5000
Neutral	6197	30	3204	0	50	5000
Sad	6076	31	2460	28	66	5000
Surprise	4001	30	1619	83	63	5000
Total	35880	213	15339	309	400	33803

The distribution of data in each class is highly imbalanced. The image resolution is varied in each dataset, so all the images are resized into standard image resolution of 128×128.

3.2 Robustness Analysis

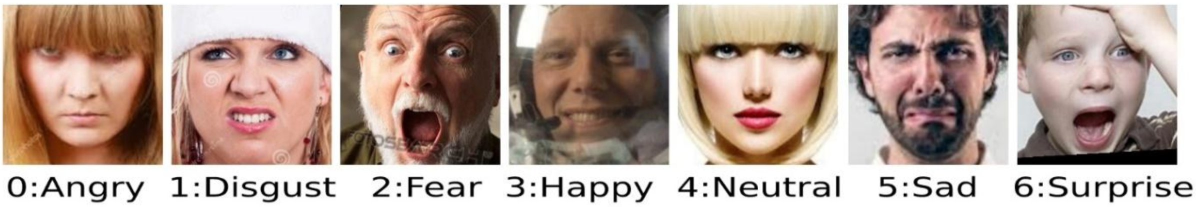

Figure 3. Influence of salient feature in facial expression

The salient features in the face that influence the way humans perceive facial expression such as the mouth, the cheek muscles, the eyebrows and size of eyes as shown in .Figure 3. Influence of salient feature in facial expression. The Dlib framework is most popular for detecting facial landmarks where 68-point landmark detectors identifies 68 points ((x,y) coordinates) in a human face as shown in Figure 4. 68 facial landmarks on face image from JAFFE These points localize the region around the eyes, eyebrows, nose, mouth, chin and jaw [35]. The studies of the faces showed that in human perception, the most important is the periocular region, followed by the mouth and then the nose [36]. The studies also suggest that the eyebrows could be even more important than the eyes due to their importance in non-verbal communication and their prominence as large, high-frequency facial features [37]. Based on the position of these features, different size rectangular regions of interest (ROI) on the face were established and masked. The occlusion operation corresponds to a simple overlay of a black rectangle over the ROI.

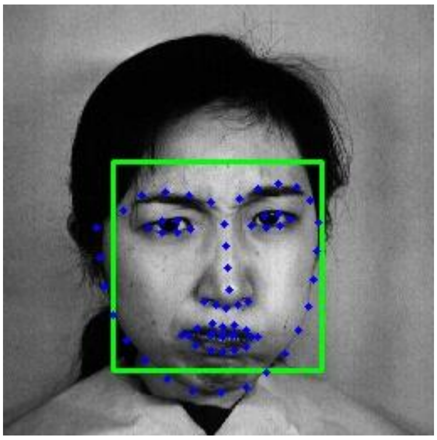

Figure 4. 68 facial landmarks on face image from JAFFE

Occlusion means conditions where parts of the critical facial features are obscured, either artificially (synthetic occlusions) or in real-world scenarios (natural occlusions). Synthetic occlusions involve deliberately modifying facial images to simulate occlusions which allows for controlled experimentation to assess how the model reacts to different types of occlusions. Therefore the mask of $8 \times 8, 16 \times 16, 24 \times 24$ pixel size is introduced randomly in the image using the Dlib framework as shown as Figure 5. Random patch (8x8,16x16,24x24), mouth and eye occluded on CK+ dataset. These masks represent unpredictable obstructions, such as random shadows, sunglasses, face mask or small objects blocking parts of the face. The mouth region is occluded using a mask to assess the ability of process model to recognize expressions without information from the mouth which is a key region for expressions like happiness or sadness. Similarly, the eye region is occluded to test the performance of model recognize the crucial for emotions such as surprise or fear.

3.3 Attention Based CNN Model

A convolutional neural network is a type of neural network commonly used for pattern recognition and classification due to its high performance in data such as image. Each layer processes a multi-dimensional array of numbers and outputs another multi-dimensional array of numbers. The proposed model integrates channel and spatial attention mechanisms with a CNN to enhance feature extraction from non-occluded portion of the image. The channel attention module computes attention weights for each RGB channel, emphasizing "what" features are meaningful where as the spatial attention module maps pixel to identify "where" important features are located.

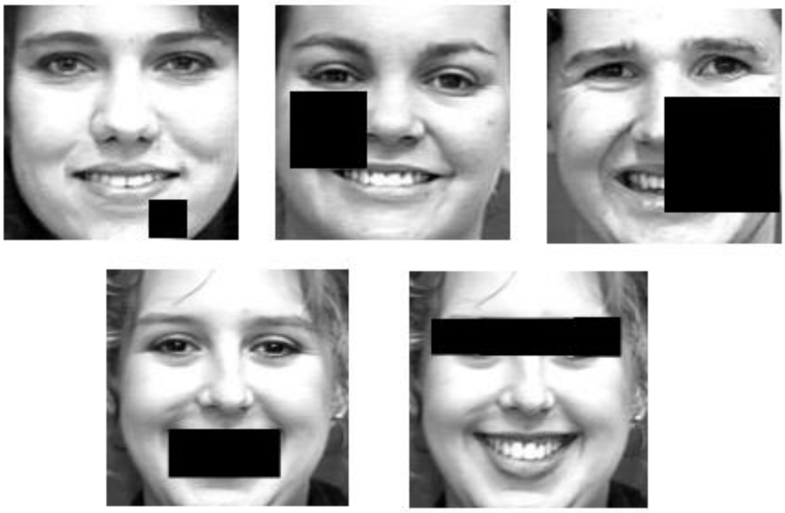

Figure 5. Random patch (8x8,16x16,24x24), mouth and eye occluded on CK+ dataset

The two modules are inserted between convolution layers in sequential manner in the proposed model. Given an input feature $\mathcal{F}: \mathcal{R}^{C\times\mathcal{H}\times\mathcal{W}}$ as input, channel module infers a 1D channel map $\mathcal{M}_c: \mathcal{R}^{C\times 1\times 1}$
and 2D spatial map $\mathcal{M}_s: \mathcal{R}^{1\times\mathcal{H}\times\mathcal{W}}$ as illustrated in Figure 6. The overview of Convolutional block attention module. The channel and spatial module can be expressed as:

$$\mathcal{F}' = M_c(\mathcal{F}) \otimes \mathcal{F} \tag{1}$$

$$\mathcal{F}'' = M_s(\mathcal{F}') \otimes \mathcal{F}' \tag{2}$$

where $\otimes$ denotes element-wise multiplication and $\mathcal{F}''$ is the final refined output.

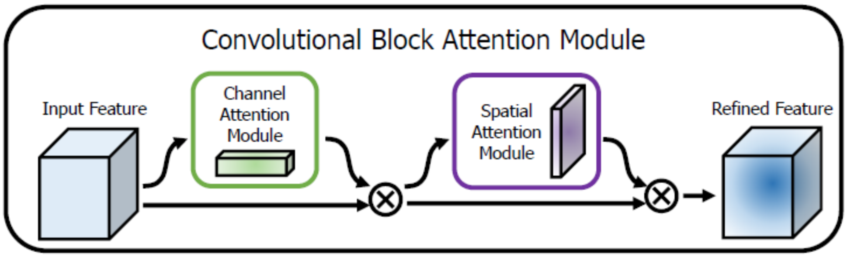

Figure 6. The overview of Convolutional block attention module

The softmax function is used to convert raw scores (logits) from a neural network into probabilities for each class. For a given i-th class, the softmax function (s_i) of the neural network is given by

$$\mathrm{f}(s_i) = \frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}} \tag{3}$$

where s_j are the scores inferred by the net for each class in $C = 7$ classes.The facial emotion recogination consists of seven categories, i.e., Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. The categorical cross-entropy loss quantifies the difference between the predicted probability distribution (from softmax) and the ground truth is given by

$$CE = -\sum_{i=1}^{C} t_i \log f(s_i) \qquad (4)$$

where t_i is the ground truth label for class i and $\log f(s_i)$ is the logarithm of the predicated probability for class i obtained from the softmax. The cross-entropy loss focuses on penalizing the model when it assigns low probability to the correct class, encouraging the model to improve its predictions for the true class.

An attention based convolutional neural network is an end-to-end model designed to imitate face recognition concentrating on critical regions of an image, such as non-occluded facial features, to make accurate predictions. The input of CNN are the pre-processed facial image, and the output is the probability of the seven categories of facial expressions as shown as Figure 7. Block diagram of attention based CNN model.

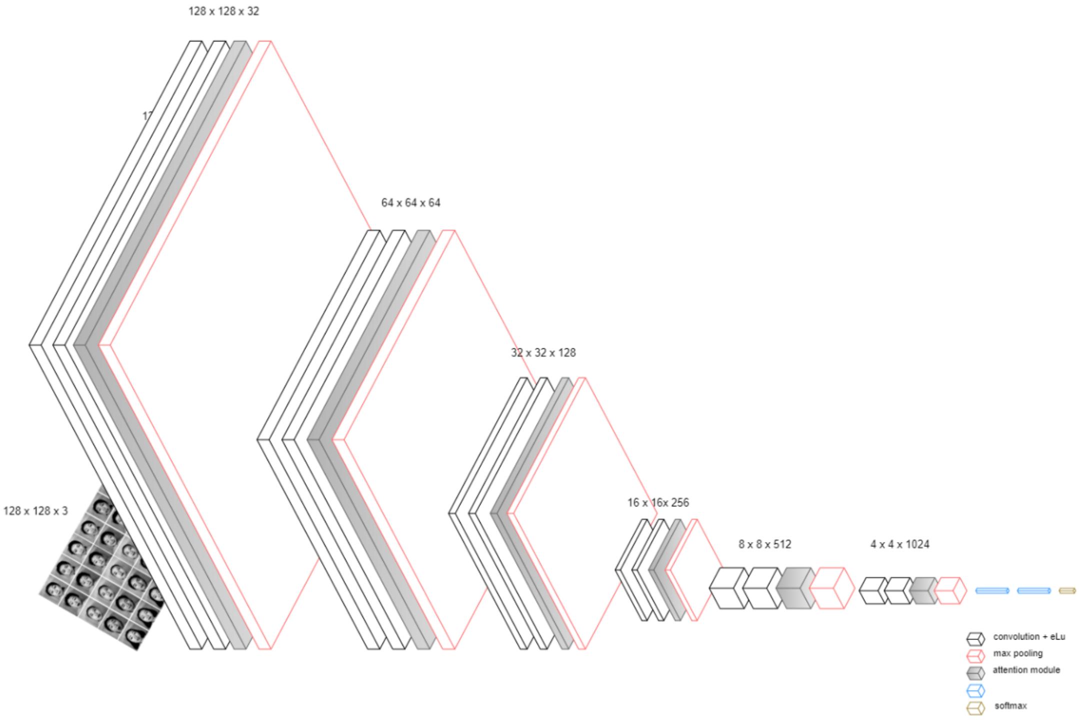

Figure 7. Block diagram of attention based CNN model

3.4 Performance Evalution

Each class of a facial emotion recognition consists of an imbalanced dataset. The evaluation metrics accuracy, precision (P), recall (R) and F1-score (F1) can be applied to evaluate the performance of the facial emotion recognition(FER) system under occluded conditions.

$$\mathrm{P} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$$

$$R = TP/(TP+FN)$$

$$F_1 = 2PR/(P+R)$$

$$accuracy = (TP + TN)\,/\,(TP + TN + FP + FN)$$

where TP represents the number of correctly classified facial expressions, FP denotes the number of misclassified expressions, including errors caused by occlusion, and FN denotes the number of actual expressions the model fails to recognize, often due to occlusion. The plotted ROC of model helps to categorize the quality of the model by showing how well a classifier performs a specific classification task.

4. Results and Discussion

The effectiveness of the model proposed in this study to recognize the facial emotion on the public dataset natural environment (Affectnet, FER2013, RAF-DB and FED-RO) and synthetic (JAFFE and CK+) under the occlusion conditions using the experimental setup as shown in Table 2.

Table 2. Experimental Setup of FER under occlusion

Parameters	**Value**
Train-test ratio	80:20
Batch size	64
Learning rate	0.001
Activation function	Elu
Optimizer	Adam
Dropout	0.2

4.1 Performance on Synthetic Occlusion Using CK+ Dataset

The CK+ lab-generated dataset used to evaluate the robustness of the proposed model under controlled occlusion scenarios. The controlled nature of this dataset allows to randomly mask with sizes of 8x8, 16x16, and 24x24 pixels and facial landmark-based occlusions of critical areas such as the eyes and mouth. The reliability of the proposed model is ensured by k-fold cross-validation where dataset is splits into 3, 5, and 10 folds which minimizes the risk of overfitting and providing a comprehensive evaluation.

Table 3. Result of k-fold cross validation on CK+ dataset

Occlusions	**Accuracy (%)**		
types	3-folds	5-folds	10-folds
8×8	86.04	89.79	92.15
16×16	88.37	86.00	84.31
24×24	84.44	80.00	77.77
Eye	97.67	93.61	94.11
Mouth	86.04	91.48	94.00

Table 3 shows the result of the k-fold cross validation on CK+ dataset under different occlusion types. The dataset was used to evaluate how well the model performs when critical facial features, which are deliberately occluded. The smaller occlusion (8×8) had a negligible impact on performance due to small portion of essential features is obscured. The spatial attention mechanism can focus on the unoccluded parts of key features and extract meaningful details where as the channel attention weighs can compensate for the limited information loss due to small occlusions. The model can rely on the symmetry of the face to infer the expression. Larger masks (24x24) more likely obscured both symmetrical counterparts, leaving the model with insufficient data, which may lead to loses of vital information required for emotion classification. Therefore, the larger masks significantly reduced performance of the model, as mask obscured more critical facial features. The spatial attention maps focus on fewer meaningful regions are available for the model to focus on, reducing its ability to compensate for the occlusion. Similarly, the channel attention can emphasize the importance of channels which cannot overcome the absence of spatial information caused by extensive masking. The occlusion of the landmark region can significantly increases the difficulty of accurately identifying the expression. Such occlusion can result in the misclassification of emotions. For instance, a smile is a defining feature of happiness. If the mouth region is occluded, the model might classify the expression as neutral or another emotion since the critical feature is hidden, particularly when the eyes do not convey strong emotions.

Table 4. Comparison with state-of-art on synthetic occluded dataset

Methods	**Eye occluded**	**Mouth occluded**
pACNN [20]	96.50	93.92
gACNN [12]	96.57	93.88
Proposed Method	97.67	94.00

The results show the effectiveness of the proposed model on the CK+ dataset outperformed current state-of-the-art under controlled occlusion scenarios as Table 4. The spatial attention mechanism likely compensates by focusing on visible areas, which are still informative for most emotions. The channel attention mechanism emphasizes important color and texture patterns, helping extract meaningful cues from the unoccluded parts of the face.

4.2 Performance on Real Occlusion

The model was trained on a combination of occlusion of realistic images from RAF-DB and AffectNet ensuring the model is exposed to diverse patterns during training. The trained model is evaluated on the FED-RO dataset. The confusion matrix for the proposed model on the FED-RO dataset revealed insights into its performance on specific emotions as shown in Figure 8. The attention mechanism of the proposed model highly classified emotions such as happy, sad, surprise of the occlusion images due to compensation made by extracting the features from visual area. However, the fear and disgust emotions are often misclassified to surprise and anger respectively due to highly correlated features. For example, both fear and surprise involve a widening of the eyes, making them visually similar. Fear may distinctive feature include slight tension in the face, such as stretched lips.

Similarly, the surprise typically features a dropped jaw or open mouth without tension. If the mouth region is occluded, the model extracts features primarily from the visual area like the eyes and eyebrows, which look similar for both emotions. The attention mechanism may not full compensate the emotions due to lack of sufficient distinctive feature and the CNN may map to neighboring regions in the feature space because the extracted features (cluster) are highly similar.

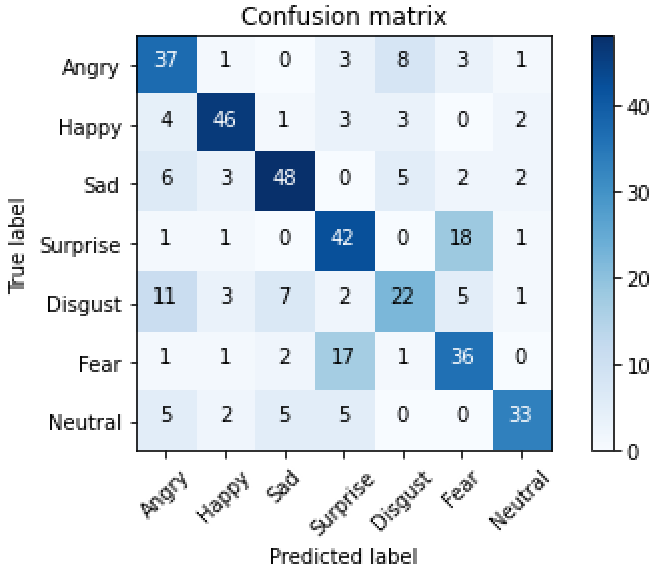

Figure 8. Confusion Matrix of FED-RO dataset using real occlusion images

Table shows the proposed model with attention mechanisms outperformed state-of-the-art methods (gACNN and pACNN) by approximately 2-3%, highlighting the effectiveness of attention in real-world scenarios. Both gACNN and pACNN models treat all parts of the input equally, including occluded areas, leading to the inclusion of irrelevant or misleading information. However, the proposed model with attention mechanisms actively suppress irrelevant occluded regions and prioritize meaningful features, making it more robust under occlusion conditions.

Table 5. Comparison with state-of-art on real occluded dataset

Method	**Accuracy**
gACNN	63.00
pACNN	64.22
Proposed without attention	62.75
Proposed with attention	66.00

The accuracy of model dropped to 62.75% when the attention mechanisms (channel and spatial attention) are removed. The channel attention mechanism can focus on visible features with distinguishing color patterns which ensures better feature extraction. This demonstrates that the attention mechanisms contribute significantly to improving the model’s ability to focus on visible and meaningful regions of the face, even under occlusion. The spatial attention shifts focus to visible features to infer the emotion, while ignoring occluded parts. Therefore, the proposed model with the attention mechanism improved robustness to occlusion by reducing the influence of the occluded areas and leverages the visible, unoccluded features for classification. However, the model without attention mechanisms processes all parts of the image equally, including occluded regions, which may introduce noise or irrelevant information.

4.3 Cross-Dataset Evaluation

Cross-dataset testing is conducted to assess the generalization of the proposed model across different datasets, where the model is trained on two datasets RAF-DB and AffectNet (real occlusion) separately, and testing was performed on clean and occluded images from the CK+ (synthetic) dataset as shown in Table 6. The gACNN performs better than the proposed model when trained on RAF-DB. The RAF-DB is a relatively smaller dataset with limited diversity and lower image quality. The simplicity of RAF-DB allows gACNN to capture relevant features effectively without needing the broader generalization capabilities of the proposed model. The proposed model employs both channel and spatial attention, which require larger and more diverse datasets (like AffectNet) to function optimally. Thus, the proposed model may overfit or fail to utilize its full potential due to limited variability in the training data.

Table 6. Cross-dataset result comparison with state-of-the-art

Methods	**RAF-DB**		**AffectNet**	
	Clean CK+	Occluded CK+	Clean CK+	Occluded CK+
gACNN	83.27%	78.05%	55.33%	52.47%
pACNN	85.07%	80.54%	58.78%	54.84%
Proposed Model	75.08%	67.96%	87.38%	82.20%

AffectNet is a much larger and more diverse dataset compared to RAF-DB. It contains a wide variety of expressions, poses, and occlusions, enabling the proposed model to learn more generalized and robust features. The channel and spatial attention mechanisms in the proposed model excel at extracting robust features from large, diverse datasets like AffectNet. which allows the model to adapt to occlusions by focusing on unoccluded regions and emphasizing relevant features. Thus, the proposed model significantly outperforms gACNN and pACNN on both unoccluded and occluded CK+. The drop in accuracy between clean and occluded CK+ images was significant but reasonable, indicating that the attention mechanism effectively mitigated the impact of occlusion.

5. Conclusions

This paper presents an attention-based CNN model for facial expression recognition under occlusion conditions. By integrating both channel and spatial attention mechanisms, the model enhances feature extraction and selectively focuses on unoccluded facial regions, improving recognition performance in obscured facial images. The attention module allows the CNN to effectively emphasize meaningful facial features while minimizing the impact of occlusions. The channel attention module identifies critical RGB features, while the spatial attention module emphasizes key areas of the face, enhancing feature extraction and improving recognition performance. The proposed model is rigorously evaluated on both synthetic and real-world occlusion datasets, including RAF, FED-RO, CK+, JAFFE, FER2013, and AffectNet, ensuring its robustness across various occlusion scenarios. Experimental results demonstrate that the model achieves 66% accuracy on occluded images, outperforming baseline methods. Additionally, cross-dataset and k-fold validation confirm the model's ability to generalize effectively across different datasets and occlusion patterns, further validating its reliability in real-world applications.

However, a limitation of the proposed model is observed, particularly in distinguishing between emotions with highly correlated features (e.g., fear vs. surprise, disgust vs. anger) under severe occlusion. To address this limitation, future work will explore the integration of multi-modal data, such as combining facial expressions with voice and body language cues to improve classification performance. Additionally, incorporating transformer-based architectures may further enhance the model's feature extraction capabilities and robustness to occlusion.

References

[1] Billah, M., Wang, X., Yu, J. and Jiang, Y. Real-time goat face recognition using convolutional neural network. Computers and Electronics in Agriculture, 2022, 194, 106730.

[2] Ikromovich, H.O. and Mamatkulovich, B.B. Facial recognition using transfer learning in the deep CNN. Open Access Repository, 2023, 4(3), 502–507.

[3] Sahan, J.M., Abbas, E.I. and Abood, Z.M. A facial recognition using a combination of a novel one-dimension deep CNN and LDA. Materials Today: Proceedings, 2023, 80, 3594–3599.

[4] Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C. and Wang, H. Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 7660–7669.

[5] Chen, Y., Wang, T., Wu, H. and Wang, Y. A fast and accurate multi-model facial expression recognition method for affective intelligent robots. In: 2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR), 2018, 319–324.

[6] Zhao, L., Wang, Z., Wang, X. and Liu, Q. Driver drowsiness detection using facial dynamic fusion information and a DBN. IET Intelligent Transportation Systems, 2018, 12(2), 127–133.

[7] Shang, Y., Yan, F., Liu, Y. and Li, Q. Facial emotion recognition based on auxiliary classifiers and SENet module. In: Fourth International Conference on Digital Signal and Computer Communications (DSCC 2024), Rashid, T.A. and Yue, Y. (Eds.), Guangzhou, China: SPIE, 2024, 10. DOI: 10.1117/12.3033244.

[8] Yang, H., Ciftci, U. and Yin, L. Facial expression recognition by de-expression residue learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 2168–2177.

[9] Zeng, J., Shan, S. and Chen, X. Facial expression recognition with inconsistently annotated datasets. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018, 222–237.

[10] Zhang, F., Zhang, T., Mao, Q. and Xu, C. Joint pose and expression modeling for facial expression recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 3359–3368.

[11] Xia, H., Li, F., Tan, Y. and Song, S. Occlusion-aware multi-scale attention consistency network for facial expression recognition. In Review, 2024. DOI: 10.21203/rs.3.rs-4740494/v1.

[12] Li, Y., Zeng, J., Shan, S. and Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 2019, 28(5), 2439–2450.

[13] Ou, W., You, X., Tao, D., Zhang, P., Tang, Y. and Zhu, Z. Robust face recognition via occlusion dictionary learning. Pattern Recognition, 2014, 47(4), 1559–1572.

[14] Li, J., Jin, K., Zhou, D., Kubota, N. and Ju, Z. Attention mechanism-based CNN for facial expression recognition. Neurocomputing, 2020, 411, 340–350.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30.

[16] Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D. and Jia, J. PSANet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018, 267–283..

[17] Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S. CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018, 3–19.

[18] Lee, D., Jang, K., Soo, Y.C., Lee, S. and Son, K. A study on the super resolution combining spatial attention and channel attention. Applied Sciences, 2023, 13(6), 3408.

[19] Sun, J., Li, X., Zhang, Y., Tang, H., Ding, Y. and Zhang, Y. Fusing spatial attention with spectral-channel attention mechanism for hyperspectral image classification via encoder–decoder networks. Remote Sensing, 2022, 14(9), 1968.

[20] Li, Y., Zeng, J., Shan, S. and Chen, X. Patch-gated CNN for occlusion-aware facial expression recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, 2209–2214.

[21] Liang, X., Meng, Y., Wang, X., Li, Y. and Zhang, J. Patch attention layer of embedding handcrafted features in CNN for facial expression recognition. Sensors, 2021, 21(3), 833. DOI: 10.3390/s21030833.

[22] Hua, Y. and Xu, X. Patch attention network for video facial expression recognition. In: Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition, Xiamen, China: ACM, 2022, 689–694. DOI: 10.1145/3573942.3574080.

[23] Gan, Y., Chen, J., Yang, Z. and Xu, L. Multiple attention network for facial expression recognition. IEEE Access, 2020, 8, 7383–7393.

[24] Zhao, X. Occluded facial expression recognition based on deep learning. Applied Computational Engineering, 2024, 95(1), 258–262. DOI: 10.54254/2755-2721/95/20241661.

[25] Lu, Y., Wang, S., Zhao, W. and Zhao, Y. WGAN-based robust occluded facial expression recognition. IEEE Access, 2019, 7, 93594–93610.

[26] Bai, Y., Chen, L., Li, M., Wu, M., Pedrycz, W. and Hirota, K. Partially occluded face expression recognition with CBAM-based residual network for teaching scene. In: 2023 China Automation Congress (CAC), Chongqing, China: IEEE, 2023, 6052–6057. DOI: 10.1109/CAC59555.2023.10450205.

[27] Pan, B., Wang, S. and Xia, B. Occluded facial expression recognition enhanced through privileged information. In: 27th ACM International Conference on Multimedia, 2019, 566–573.

[28] Hu, K., Huang, G., Yang, Y., Pun, C.-M., Ling, W.-K. and Cheng, L. Rapid facial expression recognition under part occlusion based on symmetric SURF and heterogeneous soft partition network. Multimedia Tools and Applications, 2020, 79, 30861–30881.

[29] Liu, Y.H. Feature extraction and image recognition with convolutional neural networks. Journal of Physics: Conference Series, 2018, 1087(6), 062032.

[30] Mollahosseini, A., Hasani, B. and Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 2017, 10(1), 18–31.

[31] Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A. and Bengio, Y. Challenges in representation learning: A report on three machine learning contests. In: Lee, M., Hirose, A., Hou, Z.-G. and Kil, R.M. (Eds.), Neural Information Processing, Berlin, Heidelberg: Springer, 2013, 117–124.Li, S., Deng, W. and Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 2584–2593.

[32] Lyons, M., Kamachi, M. and Gyoba, J. The Japanese female facial expression (JAFFE) dataset. Zenodo, 2019. DOI: 10.5281/zenodo.3451524.

[33] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. 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, 2010, 94–101.

[34] King, D.E. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 2009, 10(60), 1755–1758.

[35] Fraser, I.H., Craig, G.L. and Parker, D.M. Reaction time measures of feature saliency in schematic faces. Perception, 1990, 19(5), 661–673. DOI: 10.1068/p190661.

[36] Sadr, J., Jarudi, I. and Sinha, P. The role of eyebrows in face recognition. Perception, 2003, 32(3), 285–293. DOI: 10.1068/p5027.