MaskTheFER: Mask-Aware Facial Expression Recognition using Convolutional Neural Network

Cheng Jiang*, Md Rakibul Hasan†, Tom Gedeon† and Md Zakir Hossain*†
*School of Computing, Australian National University, Canberra, Australia
†School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, Australia
{Cheng.Jiang, Zakir.Hossain}@anu.edu.au, {Rakibul.Hasan, Tom.Gedeon, Zakir.Hossain1}@curtin.edu.au

***Abstract*—Facial Expression Recognition (FER) plays a crucial role in understanding people's internal states. While existing FER methods have achieved high accuracy when facial features are fully visible, the widespread use of masks during the COVID-19 pandemic has led to a significant loss of facial information. Although a few studies have explored masked FER by masking publicly available datasets, the absence of benchmark datasets on masked facial expressions poses a challenge. In this research, we address this gap by generating and publishing masked versions of two well-known datasets, namely FER2013 and CK+. Our proposed approach focuses on upper facial features in masked images to effectively handle the occlusion caused by masks. Initially, facial landmarks are detected in the masked images, which are then used to crop and align the images, retaining only the region surrounding the eyes. Subsequently, a Convolutional Neural Network (CNN) model based on a modified VGGNet architecture, incorporating fewer convolutional filters and layers, is trained and evaluated on the newly generated MaskedFER2023 and MaskedCK+ datasets. Our method achieves competitive performance – accuracies of 0.6189 and 0.6356 on the MaskedFER2023 and MaskedCK+ datasets, respectively – compared to existing state-of-the-art occlusion-aware and mask-aware FER methods. Additionally, we delve into the impact of masks on the recognition of different emotions. Our experimental results demonstrate that face masks significantly impede the recognition of certain expressions, particularly 'Sad', while other emotions like 'Surprise' exhibit lower sensitivity to masks. Implementation, experimentation and evaluation are publicly available at `https://github.com/hasan-rakibul/MaskTheFER`.**

***Index Terms*—convolutional neural network, CK+, facial expression recognition, FER2013, masked face, MaskedFER2023**

I. Introduction

Facial Expression Recognition (FER) is used to analyse individuals' facial expressions, enabling the inference of emotions such as happiness, sadness, anger and surprise. This ability to discern emotions has significant implications for improving human-machine interaction and finds diverse applications in fields such as healthcare, social media, entertainment and gaming. While many deep learning-based methods for facial expression recognition have achieved commendable accuracy when provided with complete facial information, research on recognising facial expressions from occluded faces remains limited. These occlusions can manifest in various forms, such as glasses, hair, or hands, each causing varying degrees of loss in facial information [1].

Cheng Jiang and Md Rakibul Hasan are co-first authors.

Amidst the COVID-19 pandemic, face masks have emerged as essential and effective preventive measures. Compared to other types of occlusions, face masks have a more pronounced impact, rendering the facial features below the nose bridge completely invisible. Consequently, differentiating emotions such as fear and surprise, as well as sadness and disgust, becomes more challenging, as these emotions primarily rely on the regions surrounding the mouth. Consequently, facial expression recognition from masked faces remains a challenging computer vision task.

To address this challenge, recent research has proposed several approaches that are specifically designed to handle the occlusions in images, such as Region Attention Network (RAN) [2], Attention-CNN (ACNN) [3] and Occlusion-Adaptive Deep Network (OADN) [4]. These approaches incorporate attention mechanisms to highlight relevant areas in an image while downplaying irrelevant regions. To some extent, these methods have improved the recognition accuracy of FER from masked faces. However, it is important to note that occlusion-aware is a broader concept compared to mask-aware. In the context of occlusion-aware, accurately identifying the occluded area is crucial for subsequent tasks, such as generating precise attention maps. However, this sub-task itself may be more challenging than anticipated, as accurately delineating occlusion boundaries proves more difficult than mere detection. Errors in this step can significantly impact the overall accuracy of facial expression recognition. Conversely, in the context of mask-aware, the occlusion area is confined to the region around the mouth. Consequently, in the case of FER with masked images, the model's performance can be further enhanced by effectively leveraging this additional information.

In this study, we propose a mask-aware FER pipeline, *MaskTheFER*, specifically designed to handle face masks in images. We name our pipeline *MaskTheFER* because we first mask normal images using the MaskTheFace algorithm proposed by [5], and then we build our system to perform FER on the generated dataset. The FER system consists of two branches: a cropping branch based on facial landmarks and a CNN-based classification branch. The cropping branch extracts a small region from a masked image, preserving only the eyes, eyebrows, forehead and upper nose bridge while discarding the face mask, hair and background. This step can be accomplished by employing pre-trained face landmark detectors such as `Dlib` library [6] or Multi-task Cascaded Convolutional

979-8-3503-8220-4/23/$31.00 ©2023 IEEE

DOI 10.1109/DICTA60407.2023.00069

Networks (MTCNN) [7]. As the datasets comprise images taken from varying angles or featuring faces with different tilts, the input images are initially aligned based on the positions of visible landmarks to ensure consistent cropping.

The resulting cropped images are then fed into the classification branch for predicting facial expressions. This branch is a modified version of the VGG architecture [8] known for its deep structure and utilisation of small convolutional filter kernels. To better accommodate the characteristics of small and low-resolution input images, the hyperparameters of the classification branch are fine-tuned.

The major contributions of this paper are summarised as follows:

1) The generation of visually appealing masked versions of widely used FER datasets, namely FER2013 [9] and CK+ [10]. Our generated MaskedFER2023 and MaskedCK+ datasets will be publicly available for benchmarking mask-aware FER tasks.
2) The proposal and effectiveness validation of a cropping-based image pre-processing method that selectively extracts key areas while discarding irrelevant regions.
3) The introduction of a variant of the VGG classifier, where hyperparameters such as depth, filter number and learning rate have been fine-tuned to optimise performance for the task at hand.
4) Evaluation of the proposed method on the generated MaskedFER2023 and MaskedCK+ datasets, showcasing competitive performance as compared to state-of-the-art methods on other variants of the FER2013 dataset masked by different techniques.
5) A comprehensive investigation into the impact of face masks on the recognition of the seven basic emotions – Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise – categorised in the original FER2013 and CK+ datasets.

II. Related Work

A. Facial Landmark Detection

Facial landmark detection plays a fundamental role in various face analysis tasks by identifying specific facial points such as the eyes, nose, and mouth. These landmarks find extensive applications in face recognition, emotion analysis, and other downstream tasks. In this paper, facial landmark detection is pivotal for generating synthetic face masks and cropping the eye regions.

One well-known facial landmark detection approach, introduced by Kazemi and Sullivan [6], employs an ensemble of regression trees and shape-indexed features to achieve real-time landmark detection. The widely used 68 landmark function in the Python `Dlib` library is based on this work. Conversely, Zhang *et al.* [7] proposed a deep learning-based method called Multi-task Cascaded Convolutional Networks (MTCNN). MTCNN comprises three subnetworks: Proposal Network (P-Net), Refinement Network (R-Net) and Output Network (O-Net). By formulating face landmark detection as a regression problem, MTCNN provides five key landmarks for each detected face.

It is important to note that these landmark detection methods were originally developed using datasets that did not incorporate face masks. However, empirical evidence by Dharanesh and Rattani [11] demonstrates that these methods can still yield relatively accurate results on masked images. This robustness makes face landmark detection applicable to various occlusion-aware or mask-aware FER approaches, including the approach presented in this paper.

B. Occlusion-Aware FER

Occlusion-aware FER approaches aim to address the challenge of recognising facial expressions in the presence of occlusions. These methods typically adopt a sub-region-based strategy, where the images are divided into multiple patches, allowing the model to focus on non-occluded regions while paying less attention to or disregarding the occluded patches.

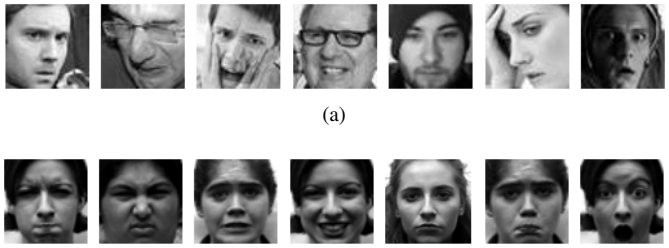
One notable occlusion-aware FER approach is ACNN [3], which is specifically designed to identify occluded patches on the face and emphasise the most informative non-occluded patches. ACNN consists of two branches: patch-based ACNN (pACNN) and global-local-based ACNN (gACNN). The pACNN branch focuses on local face patches, whereas the gACNN branch integrates both local and global features. Experimental results demonstrate that ACNN improves recognition accuracy for both non-occluded and occluded facial expressions.

Another occlusion-aware FER method, the Region Attention Network (RAN), introduced by Wang *et al.* [2], employs adaptive region attention to capture the importance of different facial regions under occlusion and various poses. Additionally, they propose a region-biased loss function that assigns higher weights to the most crucial facial regions. Similarly, the Occlusion-Adaptive Deep Network (OADN) proposed by Ding *et al.* [4] incorporates an attention mechanism through an attention map, which assigns different weights to different regions. Furthermore, to enhance robustness, OADN divides the feature maps into non-overlapping blocks, with each block independently predicting the expression, resulting in more distinctive features.

C. Mask-Aware FER

Mask-aware FER can be considered as a specific sub-problem within the broader scope of occlusion-aware FER, where the occlusion is primarily concentrated around the mouth region. Castellano *et al.* [12] proposed a method, referred to as CroppedFace, where the images are cropped to retain only the regions proximate to the eyes. Subsequently, a CNN model is employed to predict the facial expressions based on these cropped faces. This approach bears similarity to the method utilised in the present paper. However, the authors of the aforementioned work lack comprehensive data pre-processing and model optimisation, leaving room for further improvements, which are addressed in this paper.

Another two-step method, BC-AD, was introduced by Yang *et al.* [13]. In the first step, a binary classifier (BC) develops attention heatmaps that delineate the masked and unmasked



(a)

(b)

Fig. 1. Sample images from the original (a) FER2013 and (b) CK+ datasets.

areas by iterating over small patches within an image. In the second step, these heatmaps are employed to re-weight features using an attention-based deep model (AD).

Building upon BC-AD, Yang *et al.* [14] proposed an updated version called Face-mask-aware Face Parsing and Vision Transformer with a Cross-Attention Mechanism (FFP-VTC). In this method, they fine-tuned a face segmentation model, which offers improved accuracy in distinguishing between masked and unmasked regions compared to a simple binary classifier. Furthermore, the attention-based deep model (AD) was replaced with a Vision Transformer, known for its more optimal performance. Notably, their approach differs from cropping-based methods in that it does not discard any occluded regions. Instead, each region is assigned a suitable weight based on its importance.

III. MaskTheFER Pipleine

A. Masked Datasets Generation

1) Original Datasets: The FER2013 dataset (Fig. 1a) is a widely studied public dataset comprising 35,887 grayscale images with dimensions of 48×48 pixels. The dataset encompasses various facial expressions, including 4,953 images labelled as 'Angry', 547 images labelled as 'Disgust', 5,121 images labelled as 'Fear', 8,989 images labelled as 'Happy', 6,198 images labelled as 'Neutral', 6,077 images labelled as 'Sad', and 4,002 images labelled as 'Surprise'. It is important to note that the FER2013 dataset was originally collected from the internet, resulting in variations in camera angles, lighting conditions and potential inaccuracies in the annotations. A survey conducted on the Kaggle forum reported that the average human accuracy on the FER2013 dataset is approximately $65 \pm 5\%$ [9].

The Extended Cohn-Kanade Dataset (CK+) (Fig. 1b) is another frequently employed benchmark dataset for FER. It comprises 593 video sequences, with each sequence containing 10 to 60 frames, capturing the facial changes of 123 subjects. In CK+, the videos depict the transition of facial expressions from a neutral state to specific target expressions. Among these video sequences, 327 sequences from 118 subjects are labelled with seven distinct expressions, namely Angry, Disgust, Fear, Happy, Sad, Surprise and Contempt. Unlike the FER2013 dataset, the CK+ dataset is collected under controlled laboratory conditions, ensuring accurate annotations and consistent layouts across the dataset.

Although CK+ dataset may not be sufficiently large to accurately measure the capacity of models, our motivation for choosing CK+ datasets is its suitability for evaluating the impacts of face masks. To clarify, many state-of-the-art deep learning methods such as [15], [16] have achieved close to 100% accuracy on the original CK+ dataset. Their near-perfect performance suggests that the original CK+ dataset contains highly recognisable expressions with similar difficulty levels for deep learning models. Thus, the accuracy of each emotion on its masked version directly reflects the impact of face masks.

2) Masked Datasets: To date, no existing masked FER datasets have been made publicly available. Consequently, it becomes necessary to generate masked versions of existing normal FER datasets. To this end, we generate the masked version of the FER2013 dataset by leveraging MaskTheFace algorithm [5], similar to the study by Magherini *et al.* [17]. This algorithm leverages facial landmark detection, implemented using the `Dlib` library, to determine the tilt of the face and identify the key points for mask placement. With this information, the algorithm synthesises a user-specified mask, such as an N95 or surgical blue mask, on the original image at an appropriate position and orientation.

Due to various challenges, such as low resolution, incomplete faces and unusual angles in certain images of the FER2013 dataset, masking them accurately becomes difficult. We, therefore, introduced a manual threshold in the MaskTheFace algorithm. Images with a confidence level lower than this threshold are considered challenging to mask and, consequently, are discarded. This process accounts for the smaller size of the generated dataset compared to its original.

Similarly, we mask the CK+ dataset using the same MaskTheFace algorithm. In the original CK+ dataset, which consists of video frames, there can be a high correlation and similarity between adjacent frames. To address this issue and prevent significant overlap between the training and test sets, a selection process was performed. Specifically, only one frame capturing the peak of each emotion was chosen from each video sequence. Additionally, the class 'Contempt' was replaced with the class 'Neutral' to maintain consistency with the masked version of the FER2013 dataset. For the 'Neutral' class, representative images were obtained by sampling a single frame from the beginning of certain video sequences.

B. Cropping Branch

1) Motivation: As mentioned earlier, attention-based methods play a crucial role in mask-aware and occlusion-aware FER techniques. These methods generate attention maps that indicate the importance of different sub-regions in the images, allowing the model to prioritise relevant areas while disregarding irrelevant regions. However, obtaining accurate attention maps can be challenging. Among different approaches, RAN [2] utilises self-attention modulo and relation-attention modulo. It requires multiple crops of a single training image, resulting in significant computational expense. Another method, BC-AD [13], employs a binary classifier to identify masked

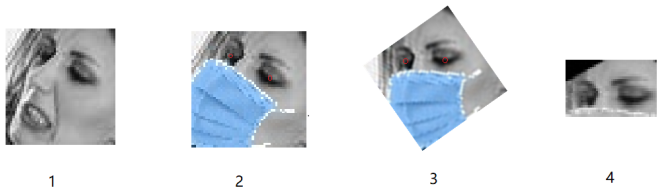


Fig. 2. Data pre-processing procedure.

and unmasked areas by iteratively 'zeroing out' square patches. The patches that most influence the classifier's prediction are deemed important unmasked areas. Although effective, this approach is not highly efficient. In FFP-VTC [14], the binary classifier is replaced with a face parser, which can segment the face into different regions, including those covered by masks. However, incorporating a face parser necessitates retraining and fine-tuning on large masked datasets, which requires additional resources, training data and time.

The mask-aware approach is more specific than occlusion-aware, as it focuses on the lower part of the face where the occlusion typically occurs. Minaee *et al.* [18] and Gan *et al.* [19] have shown that the regions near the eyes and mouth are the most informative for FER, but they are occluded when masks are worn. Based on this understanding, we propose that cropping the regions near the eyes can serve as an effective and efficient pre-processing method for FER from masked faces. By excluding the masked regions, we can eliminate the impact of irrelevant factors such as background area, mask colour and shape variations and differences in mask-wearing styles while still preserving crucial information.

2) Procedure: The entire data pre-processing procedure (Fig. 2) can be summarised as the following four steps:

Step 1: Collecting Images from the original FER2013 and CK+ datasets

Step 2: Generating MaskedFER2023 and MaskedCK+ datasets using MaskTheFace algorithm

Step 3: Horizontal alignment by the positions of eyes

Step 4: Cropping the region near the eyes so that only upper facial features are preserved

Step 3 addresses the issue of inconsistent tilts of faces, which is particularly evident in the FER2013 dataset. As depicted in Fig. 2, a regular rectangular region may not be suitable for selecting the region close to the eyes in some images. To overcome this challenge, we employ the Python `Dlib` library for face landmark detection. The face landmarks provide crucial information for determining the angle between the line connecting the two eyes and the horizontal line. Based on this angle, we rotate the image to align it appropriately. This process ensures that the selected region near the eyes is consistently oriented, regardless of variations in face tilt across the dataset.

In Step 4, the image is finally cropped to remove the background and discard the region covered by the face mask. The bounding box created by the `Dlib` face detection model determines the top, left and right borders. To determine the bottom boundary, the position of the nose is utilised, thereby excluding the region covered by the face mask. The resulting cropped region is subsequently re-scaled to a fixed size of 48×96 to serve as input to the classifier.

3) Masked and Cropped Datasets – MaskedFER2023 & MaskedCK+: The final masked and cropped versions of the FER2013 and Ck+ datasets – hereby referred to as MaskedFER2023 and MaskedCK+ datasets – contain 23,435 and 355 images, respectively. The distributions of emotions in the MaskedFER2023 and MaskedCK+ datasets are reported in Table I. We publish these generated datasets in Jiang *et al.* [20].

We split the datasets into a training set and a test set in the ratio of 85%/15% for the MaskedFER2023 dataset and 66.7%/33.3% for the MaskedCK+ dataset. The proportions of seven expressions in the training and test sets are nearly identical. Since MaskedCK+ contains only 355 images, a larger ratio for the test set is used to ensure more stable results.

To enhance the model's robustness and mitigate overfitting, data augmentation strategies are applied to the training set. However, in this study, it is important to ensure that the augmentation parameters are reasonably small to avoid discarding key features during the augmentation process, such as shifting and zooming. The relevant augmentation parameters employed in this paper are listed in Table II.

C. Classification Branch

1) Motivation: Utilising pre-trained CNN architectures as feature extractors does not offer significant advantages and may even yield inferior performance in the FER task at hand. There are several reasons for this. Firstly, the FER2013 and CK+ datasets used in this study consist of grayscale images with only one channel, whereas most pre-trained networks are trained on colour images with three channels. While it is possible to create 'fake' colour images by duplicating the grayscale channel, this process results in the loss of colour information, thereby reducing the accuracy of the extracted features.

Secondly, the datasets employed in this paper have a low resolution of 48×48, and the resolution of the cropped images prior to re-scaling may be even lower. Consequently, networks with large convolutional kernels (greater than 3×3) and deep architectures may not be suitable for this task. The utilisation of complex architectures with a high number of parameters on small and unclear images can make the optimisation process significantly more challenging.

To address these challenges, it is deemed appropriate to adopt a lightweight CNN architecture for feature extraction and train it from scratch on the generated datasets. This approach allows for better adaptation to the specific characteristics and constraints of the FER task at hand.

2) Network Architecture: We developed a CNN architecture, which is a modified version of the VGG16/VGG19 architecture and inspired by [21]. Khaireddin and Chen [21] reported state-of-the-art performance on the original FER2013

TABLE I
NUMBER OF IMAGES IN EMOTION CLASSES OF THE GENERATED MASKEDFER2023 AND MASKEDCK+ DATASETS.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
MaskedFER2023	3,048	418	3,104	6,510	4,347	3,287	2,721
MaskedCK+	45	58	25	68	50	27	82

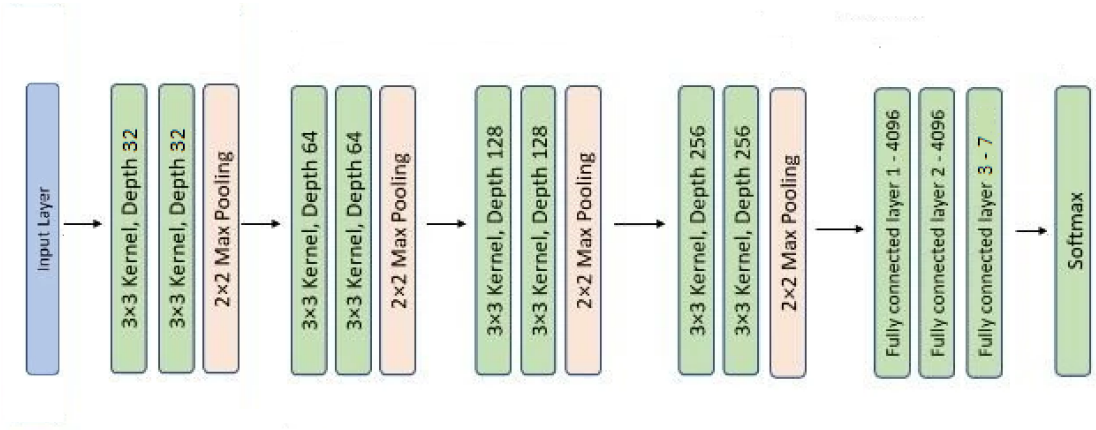


Fig. 3. The architecture of the proposed CNN model. It consists of four basic convolutional blocks, where each block contains two consecutive convolutional layers with the same number of filters, one batch normalisation layer and a 2×2 maximum pooling layer. The number of filters in each block progressively increases, starting from 32 in the first block and doubling in each subsequent block (64, 128, and 256).

TABLE II
DATA AUGMENTATION PARAMETERS.

Method	Value
Horizontal flip	True
Rotation	±15°
Zoom	±10%
Width shift	±8%
Height shift	±8%
Shear	±10%

TABLE III
IMPLEMENTATION DETAILS.

Hyperparameter	Value
Optimiser	SGD with Nesterov
Initial learning rate	0.01
Learning rate scheduler	ReduceLROnPlateau
Batch size	64 (MaskedFER2023) / 16 (MaskedCK+)
Loss function	CrossEntropy

dataset at the time of publication. To keep the network shallow and suitable for the small and unclear cropped images, we have reduced the number of filters and the depth of the architecture (two instead of three convolutional layers in the 3rd and 4th blocks) compared to the original VGG architectures.

All convolutional and linear layers (except the last linear layer) are activated by the `ReLU` activation function. The convolutional blocks are followed by two linear layers of length 4,096, which assign weights to the extracted features. To mitigate overfitting, a dropout rate of 0.30 is applied to these two linear layers. The final layer of length seven, activated by `SoftMax`, corresponds to the predictions of seven facial expressions.

Other implementation details and important hyperparameters are listed in Table III. An initial learning rate of 0.01 is dynamically reduced based on performance on the validation set (`ReduceLROnPlateau`). Specifically, the learning rate is reduced by a factor of 0.75 if the accuracy of the model does not improve after five consecutive epochs. This dynamic adjustment helps the learning rate to be high initially, enabling exploration of different areas in the parameter space, and then gradually decreases to allow for convergence to a good solution.

We implement the whole system in Python using the TensorFlow framework on a machine having an NVIDIA GeForce RTX 3070ti GPU.

IV. RESULTS AND DISCUSSION

A. Performance on MaskedFER2023

1) Overall Accuracy: The proposed method's performance on the test set is compared with three occlusion-aware methods (RAN, ACNN and OADN) and three mask-aware methods (CroppedFace, BC-AD and FFP-VTC) (Table IV). The performance of VGG19, implemented in [14], is also compared.

Our approach achieves an accuracy of 0.6189, while FFP-VTC achieves an accuracy of 0.6653 (Table IV). It is important to note that the masked versions of the FER2013 datasets used in these studies vary due to the absence of a publicly available masked version of the FER2013 dataset. Different research groups employed different techniques to mask the

TABLE IV
ACCURACY COMPARISON ON THE MASKED VERSIONS OF THE FER2013 DATASET, GENERATED USING DIFFERENT FACE-MASKING ALGORITHMS.

Reference	Face-Masking Algorithm	Method	Accuracy
[14]	FMA-3D from FaceX-Zoo [22]	VGG19	0.5148
		RAN	0.5378
		ACNN	0.5721
		OADN	0.5911
		FFP-VTC	0.6653
	CroppedFace [12]	CNN	0.5428
	AWFM [23]	BC-AD	0.6179
Proposed	MaskTheFace [5]	MaskTheFER	0.6189

faces. To ensure future research comparability, we make the MaskedFER2023 dataset publicly available.

Our proposed method slightly outperforms the BC-AD approach, which achieves an accuracy of 0.6179. The BC-AD method employs binary attention maps that resemble cropping. This similarity might explain the close performance between the BC-AD approach and our proposed method.

CroppedFace, a cropping-based method similar to our proposed approach, utilises cropped regions near the eyes as input to the CNN model. Our proposed method significantly outperforms CroppedFace (0.6189 vs 0.5428). We attribute this performance gap to two main reasons. Firstly, our proposed method incorporates additional measures in the cropping process, such as horizontal alignment of eye positions to ensure consistency between cropped images (see Fig. 2). Without this alignment, many cropped images appear visually unappealing due to varying tilts of the faces. This alignment technique improves our overall accuracy by 1.5%–2%. Secondly, the backbone architecture of CroppedFace is the classic VGG16, which may not be the most suitable for this specific task, as discussed earlier. In our paper, we carefully modified the width and depth of our backbone to enhance its adaptability to the input. Additionally, we conducted thorough hyper-parameter tuning (e.g., optimiser, learning rate, batch size) to optimise the overall performance of our model.

2) Evaluation of Different Emotion Classes: We also investigated the performance of our method on the seven emotion classes from the MaskedFER2023 dataset, as shown in Fig. 4. We compared performances with other state-of-the-art methods from two different perspectives (Table V). Firstly, the differences among mask-aware methods on different masked versions of the FER2013 datasets are studied to demonstrate the capacity of our model. Secondly, the differences in performance with the original FER2013 dataset are studied to understand the impacts of face masks.

The proposed method outperforms CroppedFace for all emotions except 'Fear'. It achieved significantly higher accuracy for emotions such as 'Disgust' (0.68 vs 0), 'Happy' (0.76 vs 0.58) and 'Neutral' (0.60 vs 0.49). It is noticeable that the CroppedFace misclassified all 'Disgust' images. This result could potentially be attributed to the imbalanced data distribution in FER2013. As mentioned earlier, there are only

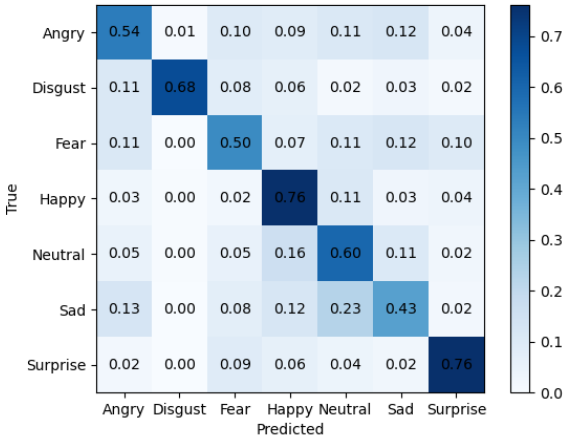


Fig. 4. Confusion matrix on MaskedFER2023 using our proposed method.

547 'disgust' images out of a total of 35,887 images in the FER2013 dataset. Compared to FFP-VTC, the proposed method achieves higher accuracy for 'Disgust' (0.68 vs 0.34), 'Happy' (0.76 vs 0.72) and 'Surprise' (0.76 vs 0.70).

To analyse the impacts of face masks on the recognition of individual emotion classes, Khaireddin and Chen [21]'s implementation is chosen as a benchmark on FER2013 because their model shares a similar backbone with ours (both are VGG based). Their model achieved a state-of-the-art accuracy of 73.28% on the original FER2013 when it was published. It is clear from Table V that 'Happy' and 'Surprise' are the two expressions that are the easiest to classify on both the original and masked datasets. Although the accuracy of these two expressions also decreases by about 10% after being masked, they are still significantly higher than other emotions on both two datasets. It is important to note that a drop of about 10% is not particularly obvious because they have a large baseline accuracy (greater than 85%). Interestingly, the accuracy of 'Disgust' on MaskedFER2023 is exceptionally higher than that on the original dataset. This may be attributed to the limited sample size for the 'Disgust' emotion in the FER2013 dataset.

In contrast, it is obvious that 'Sad' is the expression that is the most sensitive to face masks. Its accuracy drops from 0.65 to 0.43. This is probably because the loss of the lower facial features near the mouth makes it difficult for the model to distinguish it from 'Neutral'. From the confusion matrix in Fig. 4, 23% of the 'Sad' samples are misclassified as 'Neutral'. And this is also the most frequent misclassification on MaskedFER2023 for the proposed model. Similarly, the accuracy of 'Angry' also suffers a significant drop of more than 10% on the MaskedFER2023 dataset, from 0.66 to 0.54.

This result is in line with previous studies: [1] proved that 'Disgust', 'Sad' and 'Angry' are the three expressions most sensitive to masking in FER problems by experimenting on a sample of 790 participants. Furthermore, Magherini *et al.* [17] found 'Sad' to be the only misclassified emotion when

TABLE V
ACCURACY OF EMOTION CLASSES ON DIFFERENT MASKED VERSIONS OF THE FER2013 AND THE ORIGINAL FER2013 DATASETS.

Dataset	Models	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Masked versions of FER2013	CroppedFace [12]	0.46	0	0.59	0.58	0.49	0.41	0.68
	FFP-VTC [14]	0.66	0.34	0.66	0.72	0.67	0.59	0.70
	MaskTheFER (Proposed)	0.54	0.68	0.50	0.76	0.60	0.43	0.76
FER2013	VGGNet [21]	0.66	0.64	0.57	0.89	0.71	0.65	0.85

TABLE VI
ACCURACY COMPARISON ON THE MASKED VERSIONS OF THE CK+ DATASET.

Reference	Method	Accuracy
[14]	VGG19	0.4712
	RAN	0.5463
	ACNN	0.5401
	OADN	0.5362
	CroppedFace	0.5079
	BC-AD	0.5713
	FFP-VTC	0.6108
Proposed	MaskTheFER	0.6356

conducting cross-validation tests on the AffectNet dataset [24].

B. Performance on MaskedCK+

1) Overall Accuracy: The proposed model achieved an overall accuracy of 0.6356 on the MaskedCK+ dataset, which outperforms the state-of-the-art results on other masked versions of the CK+ dataset (Table VI).

Compared to the performance on the MaskedFER2023 dataset, our method demonstrates only a modest improvement of 1.67% in overall accuracy on MaskedCK+, despite CK+ containing more accurate annotations and less noise than FER2013. In our analysis, we identify two potential reasons for this outcome. Firstly, the limited size of the CK+ dataset presents challenges during the training process, which can impact the model's ability to generalise effectively. Secondly, the cropping branch in the model effectively eliminates the inconsistency of image style, specifically the variations in face tilt and position in the FER2013 dataset, which improved the alignment and uniformity of images. On the other hand, CK+ does not experience the same benefit since the majority of its images are already consistent from the beginning.

2) Evaluation of Different Emotion Classes: The recognition accuracy of each expression is shown in Fig. 5. On this small test set, the model correctly classified all 'Neutral' images. The accuracy for the 'Surprise' image is 0.89, which ranks second. It is followed by 'Happy' with an accuracy of 0.70. It can be seen that the easiest emotions to classify in the MaskedCK+ dataset align with the MaskedFER2023 datasets. Both 'Happy' and 'Surprise' maintain an accuracy of over 70% on these two datasets.

In contrast, 'Fear', 'Sad' and 'Angry' are the most difficult expressions to classify, which is also consistent with the results from the MaskedFER2023 dataset. In our opinion, this is probably because the most distinctive features for these

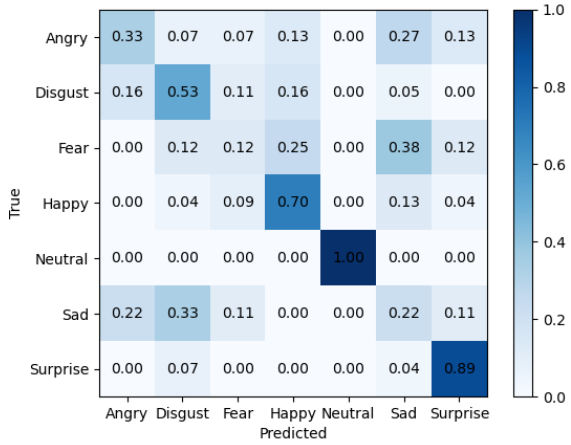


Fig. 5. Confusion matrix on MaskedCK+ using our proposed method.

Fig. 6. Some incorrect predictions by our model. In the top row, 'Sad' is misclassified as 'Fear', while in the second row, 'Angry' is misclassi fied as 'Happy'. The first, second and third columns are the original images, the masked images and the images after pre-processing, respectively.

emotions are in the lower part of the face near the mouth, which is not visible when a mask is worn.

It is worthwhile to note that a higher accuracy achieved by any certain emotion might be attributed to the fact that this emotion is inherently easier to classify (e.g., fewer outliers and less noise for this class in the original dataset) rather than a relatively minor impact from face masks.

C. Misclassification Caused By Face Mask

Fig. 6 shows two examples of how face masks lead to misclassification. These two examples indicate that the absence of lower facial features can make some expressions

unrecognisable. It further explains why the overall recognition accuracy on CK+ dropped from almost 100% to no more than 70% after the images were masked.

V. Conclusion and Future Work

In this paper, we present a mask-aware method to specifically handle occlusion in emotion recognition tasks from masked faces. The proposed method consists of a cropping-based image pre-processing branch and a CNN-based classification branch. The pre-processing branch uses pre-trained face landmark detectors to crop the image so that only the region near the eyes, which is the most informative area in masked images, is preserved. The CNN classifier is adapted from the classical VGG architecture with its depth, width and other hyperparameters fine-tuned for the task at hand.

Two masked datasets – MaskedFER2023 and MaskedCK+, with seven basic emotions – are generated and made available from their original version to evaluate the model and understand the impact of masks on FER tasks. Experimental results indicate that our proposed *MaskTheFER* pipeline achieves competitive performance on different masked versions of FER2013 and CK+ datasets compared to other state-of-the-art methods. Furthermore, we show that face masks make the recognition of some expressions, especially 'Sad', much more difficult, according to the statistics for each of the emotions. While some expressions such as 'Surprise' and 'Happy' are not sensitive, maintaining a high recognition accuracy with or without the presence of masks.

The quality of the generated masked images can be evaluated, which we leave for future work. In addition, incorporating our system with a mask-detection model, such as [25], would be another interesting avenue for future research. This integration would allow the system to be applied in real-world scenarios where both masked and unmasked faces are present.

References

[1] M. Grahlow, C. I. Rupp, and B. Derntl, "The impact of face masks on emotion recognition performance and perception of threat," *PLoS One*, vol. 17, no. 2, e0262840, 2022.

[2] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020. DOI: 10.1109/TIP.2019.2956143.

[3] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, 2019. DOI: 10.1109/TIP.2018.2886767.

[4] H. Ding, P. Zhou, and R. Chellappa, "Occlusion-adaptive deep network for robust facial expression recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2020, pp. 1–9.

[5] A. Anwar and A. Raychowdhury, *Masked face recognition for secure authentication*, 2020. arXiv: 2008.11104 [cs.CV].

[6] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, IEEE Computer Society, 2014, pp. 1867–1874. DOI: 10.1109/CVPR.2014.241.

[7] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," 2016. arXiv: 1604.02878.

[8] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, Kuala Lumpur, Malaysia, November 3-6, 2015*, IEEE, 2015, pp. 730–734. DOI: 10.1109/ACPR.2015.7486599.

[9] I. J. Goodfellow, D. Erhan, P. L. Carrier, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing - 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III*, M. Lee, A. Hirose, Z. Hou, and R. M. Kil, Eds., ser. Lecture Notes in Computer Science, vol. 8228, Springer, 2013, pp. 117–124. DOI: 10.1007/978-3-642-42051-1_16.

[10] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. A. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2010, San Francisco, CA, USA, 13-18 June, 2010*, IEEE Computer Society, 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[11] S. Dharanesh and A. Rattani, "Post-covid-19 mask-aware face recognition system," in *2021 IEEE International Symposium on Technologies for Homeland Security (HST)*, IEEE, 2021, pp. 1–7.

[12] G. Castellano, B. D. Carolis, and N. Macchiarulo, "Automatic emotion recognition from facial expressions when wearing a mask," in *Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter, CHItaly '21, Bozen-Bolzano, Italy, and online (www), July 11-13, 2021*, A. D. Angeli, L. Chittaro, R. Gennari, *et al.*, Eds., ACM, 2021, 22:1–22:5. DOI: 10.1145/3464385.3464730.

[13] B. Yang, J. Wu, and G. Hattori, "Face mask aware robust facial expression recognition during the covid-19 pandemic," in *2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021*, IEEE, 2021, pp. 240–244. DOI: 10.1109/ICIP42928.2021.9506047.

[14] B. Yang, J. Wu, K. Ikeda, *et al.*, "Face-mask-aware facial expression recognition based on face parsing and vision transformer," *Pattern Recognit. Lett.*, vol. 164, pp. 173–182, 2022. DOI: 10.1016/j.patrec.2022.11.004.

[15] M. Aouayeb, W. Hamidouche, C. Soladié, K. Kpalma, and R. Séguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," 2021. arXiv: 2107.03107.

[16] D. Meng, X. Peng, K. Wang, and Y. Qiao, *Frame attention networks for facial expression recognition in videos*, 2019. arXiv: 1907.00193 [cs.CV].

[17] R. Magherini, E. Mussi, M. Servi, and Y. Volpe, "Emotion recognition in the times of COVID19: coping with face masks," *Intell. Syst. Appl.*, vol. 15, p. 200 094, 2022. DOI: 10.1016/j.iswa.2022.200094.

[18] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021. DOI: 10.3390/s21093046.

[19] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, 2020. DOI: 10.1109/ACCESS.2020.2963913.

[20] C. Jiang, M. R. Hasan, T. Gedeon, and M. Z. Hossain, *Masked-FER2023 and MaskedCK+: Face-masked FER2013 and CK+ datasets for mask-aware facial expression recognition*, Mendeley Data, V1, 2023. DOI: 10.17632/sp3xssmzbg.1.

[21] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," 2021. arXiv: 2105.03588.

[22] J. Wang, Y. Liu, Y. Hu, H. Shi, and T. Mei, "Facex-zoo: A pytorch toolbox for face recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3779–3782. DOI: 10.1145/3474085.3478324.

[23] B. Yang, J. Wu, and G. Hattori, "Facial expression recognition with the advent of face masks," in *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*, 2020, pp. 335–337. DOI: 10.1145/3428361.3432075.

[24] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," 2017. arXiv: 1708.03985.

[25] B. Yang, M. Z. Hossain, and S. Rahman, "SS-faster-RCNN: A domain adaptation-based method to detect whether people wear masks correctly," in *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8.