Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Face-mask-aware Facial Expression Recognition based on Face Parsing and Vision Transformer

Bo Yang, Jianming Wu, Kazushi Ikeda, Gen Hattori, Masaru Sugano, Yusuke Iwasawa, Yutaka Matsuo

* KDDI Research, Inc., 2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502, Japan
b The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8654 Japan

** Article Info

Article history:
Received 14 May 2022
Revised 5 October 2022
Accepted 4 November 2022
Available online 9 November 2022

Edited by Maria De Marsico

MSC:
4A05
4A10
65D05
65D17

Facial expression recognition
Face mask
Deep learning
Vision Transformer
Face parsing
Covid-19

** Abstract

As wearing face masks is becoming an embedded practice due to the COVID-19 pandemic, facial expression recognition (FER) that takes face masks into account is now a problem that needs to be solved. In this paper, we propose a face parsing and vision Transformer-based method to improve the accuracy of face-mask-aware FER. First, in order to improve the precision of distinguishing the unobstructed facial region as well as those parts of the face covered by a mask, we re-train a face-mask-aware face parsing model, based on the existing face parsing dataset automatically relabeled with a face mask and pixel label. Second, we propose a vision Transformer with a cross attention mechanism-based FER classifier, capable of taking both occluded and non-occluded facial regions into account and reweigh these two parts automatically to get the best facial expression recognition performance. The proposed method outperforms existing state-of-the-art face-mask-aware FER methods, as well as other occlusion-aware FER methods, on two datasets that contain three kinds of emotions (M-LFW-FER and M-KDDI-FER datasets) and two datasets that contain seven kinds of emotions (M-FER-2013 and M-CK+ datasets).

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, thanks to the deep learning boom in the computer vision field, the facial expression recognition (FER) technology, which aims to deepen the understanding of human beings and human-computer interaction, has already achieved reasonable accuracy [26]. However, the partial occlusion problem remains as one of the major challenges for accurate FER, and it can be classified into two categories [54]: 1. facial accessories, such as sunglasses, scarves, and masks. 2. random objects in the facial area, such as hairs, cups, and glasses. Wearing a face mask, which could be classified as a facial accessory, has not been extensively studied compared with other occlusion situations. The COVID-19 has spread to every part of the world during the nearly 3-year period since 2019, and although many people have been vaccinated multiple times, wearing face masks is still considered by many organizations and governments to be an effective means of preventing transmission of the coronavirus. However, psychologists also argue that the prolonged use of face masks is making it very difficult for counterparts to read the emotions of the wearer, as well as affecting social interaction between individuals [3]. Thus, solving the problem of face-mask-aware FER is becoming increasingly important in the interactions between people, as well as between people and machines. When half of the face is covered, including the region around the mouth and nose that conveys a lot of non-verbal information, it becomes even harder to distinguish the emotions of fear and surprise, as well as emotions of sadness and disgust [43].

In our earlier study [50], a two-stage attention deep network was proposed to address the issue of the face mask problem in FER in relation to three emotions. The two stages are as follows: 1. the masked/unmasked binary classification stage to determine the face mask region. 2. the masked facial expression recognition stage for face-mask-aware FER. In this paper, we present a deep learning pipeline based on face parsing and a vision Transformer with a cross-attention mechanism to improve the performance of face-mask-aware FER.

https://doi.org/10.1016/j.patrec.2022.11.004
0167-8655/© 2022 Elsevier B.V. All rights reserved.

 Corresponding author.
E-mail addresses: bo-yang@kddi-research.jp (B. Yang), ke-go@kddi.com (J. Wu), k2-ikeda@kddi.com (K. Ikeda), ge-hattori@kddi.com (G. Hattori), masugano@kddi.com (M. Sugano), iwasa@weblab.t.u-tokyo.ac.jp (Y. Iwasawa), matsuo@weblab.t.u-tokyo.ac.jp (Y. Matsuo).
The main contributions of this work can be summarized as follows:

1. To the best of our knowledge, this could be considered to be the first study that takes into account the face mask in face parsing. This could benefit other face-mask-aware downstream tasks, such as face recognition, engagement intensity prediction, and face modifications.

2. The obtained face-mask-aware face parsing model is then utilized to distinguish the unobstructed facial region as well as the face mask region from the other regions in a specific image. This approach is more precise than that of our previous study and thus contributes to improving the detection accuracy of facial expression recognition.

3. Compared to the previous study where a specific percentage (0.1 and 0.9) was used empirically, the vision Transformer-based model is used to tackle the face-mask-aware facial expression recognition task, while a cross-attention mechanism is adopted to deal with the self-adaptive re-weight of attention on the masked region and unmasked region.

4. The proposed approach outperforms existing state-of-the-art face-mask-aware FER methods as well as occlusion-aware FER methods in dealing with face mask problems in FER. The proposed method is evaluated not only on our private datasets M-KDDI-FER and M-FER-T, but also on M-LFW-FER with three kinds of emotions, and M-FER-2013 along with M-CK+ with seven kinds of emotions, which are automatically processed as facial image wearing face masks based on publicly available datasets LFW-FER, FER-2013, and CK+.

2. Related works

2.1. Occlusion problem in computer vision and FER

Although FER based on deep learning has achieved some degree of success, performance is seriously degraded if the face is significantly occluded. Generally, the approaches used to deal with occlusion problems in computer vision can be classified into three main categories: reconstruction-based, holistic-based, and sub-region-based approaches. The majority of existing FER research has focused on evaluating approaches in relation to non-occluded facial images [20,38,55–57]. However, there is also an increasing number of studies focusing on partial occlusion problems in two-dimensional FER and three-dimensional FER [11,47,60].

The reconstruction-based approaches for occlusion-aware FER first restore the occluded facial regions using methods such as GAN (generative adversarial network) and then treat it as if it were a non-occluded FER. These approaches may work when only a few parts of the face are obscured [31]. However, the temptation to restore the facial region occluded by the face mask is not so much as impractical but more like a gamble, for more than half of the face is completely obscured.

The holistic-based approaches take the position that sparse signal representation is the key to object recognition and is robust to occlusion. These approaches have been reported by [46] and [39]. Although these studies were able to demonstrate good performance and robustness in occluded object classification tasks and face recognition, occluded FER was not evaluated.

The sub-region-based approaches, one of the most common methods for occlusion-aware FER, divide the facial image into small-sized patches, regardless of whether they overlap or not. Deep learning models and attention mechanisms are usually applied to guide the model to pay more attention to non-occluded patches that are essential for classification accuracy, but ignore the occluded patches or afford them less attention. The ACNN framework for occlusion-aware FER consists of two layers: patch-based ACNN(pACNN) and global-local-based ACNN (gACNN) [27]. An attention net is proposed to eliminate the effect of occlusion, through re-weighting local representations at the patch-level and the global representations at the image level. Another study of the RAN framework presented two modules: the self-attention module and the relation-attention module [45]. This approach can adaptively capture the facial regions that are more important for accurate FER. However, the aforementioned two frameworks may fail to accurately locate large non-occluded facial regions as self-attention-based methods lack sufficient supervision. The author who proposed the OADN framework argued that the approach utilizes two branches, a landmark-guided attention branch, and a facial region branch, to locate large non-occluded facial regions [9].

The aforementioned existing occlusion-aware FER approaches based on deep learning train a model to pay more attention to non-occluded facial regions. While the key to solving the occlusion problem is modeling the occlusion, this is very difficult because occlusion is normally irregular, random, and complex [59].

2.2. Face-mask-aware FER

Face-mask-aware FER can be treated as a form of occlusion-aware FER. It seems wiser to seek solutions by focusing on non-occluded facial regions, considering that face masks are relatively more regular in shape and size compared with other occlusions. To
the best of our knowledge, only a few studies have addressed the face-mask-aware FER problem.

An existing study on FER that took face masks into account involved recognizing emotions only from the eyes region. This approach was evaluated on a masked FER-2013 dataset that included seven emotions [4]. However, other important facial areas such as the forehead are discarded when only the eyes region is isolated using landmark detection methods, which decreases the accuracy of the FER system. We have also found that the existing tools for landmark detection, such as Openface2.0 [2] used in current studies, are not accurate when the facial image includes a face mask and thus further decreases the detection accuracy of the FER system. Although some studies have addressed the problem of landmark detection robustness [16,24,48,58], there are still some issues that need to be addressed: 1. there is still a gap in detection accuracy of landmarks between a masked face and a face without occlusion. 2. utilizing the landmarks could only roughly separate the covered facial region and the uncovered image region.

In our earlier study, a two-stage attention deep network which can deal with three emotions (positive, negative, and neutral) was proposed to address the issue of the face mask problems in FER [50]. In the masked/unmasked binary classification stage, the attention mechanism is incorporated into the classifier to generate attention heatmaps of the masked region and reverse attention heatmaps of the unmasked region. In the masked FER classifier stage, the attention mechanism is also used to guide the model to focus on the facial parts most important to the FER classification results and, meanwhile, pay more attention to the unmasked region but less to the masked region. Although this approach demonstrated outstanding performance compared with other partial occlusion FER approaches, there are still two deficiencies that need to be solved: 1. using attention mechanism, only the face mask region is separated from other image regions with a relatively high accuracy. 2. the re-weight of the detected masked region and un-masked region is not adaptive. Besides, another shortcoming of the earlier work was that only three emotions were evaluated.

2.3. Face parsing for FER

Face parsing, which is a specific task in semantic segmentation, assigns pixel-level labels to each semantic component, such as the skin, mouth, and lips [44]. This could benefit many downstream tasks, such as face recognition [7,41], face synthesis [14,61], and facial attribute recognition [21,22].

Only a few studies could be found on face parsing related to facial expression. CAFP-GAN, which was introduced to advance the expression synthesis domain with face parsing transformation, takes advantage of both the knowledge of facial semantic regions and controllable expression signals [32]. Another study proposed a deep learning model based on the components of face parsing for facial expression recognition [35]. However, the “face parsing” mentioned in this paper roughly divides the face region into patches, which is different from the face parsing that assigns pixel-level semantic labels. Through a careful investigation, we found that there is still ample scope and potential to use face parsing to benefit the study of facial expression.

3. Proposed methodology

3.1. Face-mask-aware face parsing (FFP)

Image cropping, where a portion of the face is cropped from the original image, is a typical pre-treatment stage in FER [1]. As eliminating unimportant parts or occlusions such as background or a face mask has been shown to increase the accuracy of facial expression recognition [4,18,50], we are trying to locate the scope of unmasked facial regions utilizing a face parsing approach, the performance and the accuracy of which are better than the previously mentioned methods for it classifies the facial image semantic at the pixel level. Because there is no publicly available research or tools about the face-mask-aware face parsing, we would like to add a mask label to the existing face parsing dataset to re-train the face parsing model, which is illustrated in Fig. 1.

First, we utilize the automatic wearing face mask (AWFM) method to superimpose a mask on images from the CelebAMask-HQ [25] and ibugmask datasets [28], which also take into account the effect of face orientation. Several AWFM methods were proposed in 2021. A 3D model-based approach achieved a good performance in fitting masks to faces in the wild and was evaluated using a face recognition task [15]. Our previously proposed AWFM method also performed well on the facial expression recognition task, which also considers face orientation [49]. In this study, we employed the publicly available FMA-3D method from the PyTorch toolbox called FaceX-Zoo [19] to make this work more reproducible. For both face parsing datasets, the FMA-3D method is used to fit eight types of masks on the face, and eight processed facial images wearing a face mask are generated from a single original image. Meanwhile, we add the label file mask.png for the face mask and delete the label files of mouth.png, up-lip.png, and lower-lip.png, accordingly.

Further, using the obtained face-mask-aware CelebAMask-HQ and ibugmask datasets, we conduct fine-tuning on the pre-trained face parsing model to obtain a face-mask-aware face parsing model. The obtained model could be later used to process any real facial image with a face mask to get the labels for each facial part: background, skin, mask, hair, eye, eyebrow, etc., but in particular, face masks and skin, which represent the face mask region and face region including the face mask region, are processed separately. The colored label map in Fig. 1 shows the effect of the obtained model in processing a facial image with a face mask. As will be discussed in the experimental study of the FFP model, although utilizing the FFP could improve the division accuracy of occluded and non-occluded facial regions compared with previous works, it is still impossible to separate the two regions one hundred percentage precisely. Thus, it is wise to consider both regions during the next face recognition stage but pay more attention to the non-occluded facial region rather than the occluded facial region.

3.2. VIT And cross-attention-based FER (VTC-FER)

Transformer [30], as a self-attention-based structure, first demonstrated its effectiveness in the field of natural language processing, and subsequently in the computer vision field by matching or even surpassing CNNs for image classification and object classification tasks, etc. Inspired by the first study to apply vision Transformers to FER [36], we would also like to take advantage of the multi-layer Transformer and be the first to tackle the face-mask-aware facial expression recognition task. Furthermore, unilateral cross-attention is proposed to re-weight attention on the masked and unmasked facial regions to achieve the optimal performance of the proposed model [5]. To this end, as is illustrated in Fig. 2, we propose a vision Transformer and cross-attention mechanism-based FER pipeline VTC-FER, along with a face-mask-aware face parsing method to locate the unshielded facial regions.

Facial region segmentation. For a given facial image I, the size of the image matrix can be expressed as H × W × C, where H and W stand respectively for height and weight and C stands for channel number. For an RGB image, the channel number is 3, while, for a grayscale image the channel number is 1. For the purpose of unified processing, the grayscale image matrix is copied 3 times to make the channel number 3. Then, through deploying the obtained face-mask-aware face parsing model, each pixel of the im-
age \( I \) is predicted with a label. We give 0 to pixels with the label of "mask" and 1 to other pixels to generate a binary mask map \( M_m \) for the face mask region. The same manipulation is repeated to form a binary mask map \( M_s \) for the skin region, which includes the masked facial parts and unmasked facial components. It is apparent that the matrix size of \( M_m \) and \( M_s \) should be \( H \times W \times 1 \). Further, \( M_m \) is subtracted from an identity matrix with the same size of \( H \times W \times 1 \) to get a reverse binary mask map \( M_r \). To extract the binary mask map of unmasked face \( M_m \), \( M_r \) is multiplied by \( M_s \) element-wisely, using the following equation:

\[
M_u = M_r \odot M_s = M_s \odot (1 - M_m)
\]  

(1)

where \( \odot \) is the element-wise product.

**Multi-layer vision Transformer encoder.** As is well known that it is almost impossible to train a Transformer from scratch without a huge dataset, we also use a pre-trained ResNet50 model [13] to extract feature maps from the original facial image, referring to [36] and [9]. The backbone of ResNet50, which is pre-trained on ImageNet but without an average pooling layer and a fully connected layer, is employed as the feature extractor. The given facial image \( I \) is manipulated into two branches: for the face mask branch, \( I \) is multiplied by \( M_m \) element-wisely, while \( I \) is multiplied by \( M_s \) element-wisely for the unmasked face branch. The processed images of the two branches are input into the pre-trained ResNet50 model to extract feature maps \( F_m \) and \( F_s \) with the size of \((H/R) \times (W/R)\), where \( R \) is the downsampling rate of the ResNet50. The values of \( H \), \( W \) and \( R \) in this paper are set to \( H = W = 224 \) and \( R = 16 \).

A traditional vision Transformer (ViT) divides an image with specific patch sizes, projects each patch into tokens, and flattens them into a 1D sequence of patch tokens [10]. In this paper, we flattened the 2D feature maps \( F_m \) and \( F_s \) into two 1D sequences \( S_m^p \) and \( S_s^p \). For each sequence, position embedding is added to each token to obtain \( S_m^0 \) and \( S_s^0 \) using the following equations:

\[
S_m^0 = \left[ S_m^{cl} \left\| S_m^{patch} \right. \right] + S_m^{pos}
S_s^0 = \left[ S_s^{patch} \right] + S_s^{pos}
\]  

(2)

where \( S_m^{cl} \in \mathbb{R}^{1 \times C_f} \) is the CLS token for the masked face branch. \( S_m^{patch} \in \mathbb{R}^{N \times C_f} \) and \( S_s^{patch} \in \mathbb{R}^{N \times C_f} \) are the feature patches for two branches, respectively. \( S_m^{pos} \in \mathbb{R}^{(N+1) \times C_f} \) and \( S_s^{pos} \in \mathbb{R}^{N \times C_f} \) are the position embedding for two branches, respectively. \( C_f \) is the dimension of the feature output, while \( N \) is the number of feature patches.

Then the \( S_m^0 \) and \( S_s^0 \) are fed into two branches of the multi-layer Transformer encoder, respectively. For either sequence, the weight of embedding \( S \) is calculated by single-head global self-attention (SHSA) and then calculated by multi-head self-attention (MHSAs) using the following equations:

\[
\text{head}_i = \text{softmax}(Q_i K_i^T / \sqrt{d}) V_i
\]

\[
\text{MHSAs}(S) = \text{concat} (\text{head}_1, \ldots, \text{head}_n) W
\]

(3)

where \( Q_i = SW_i \), \( K_i = SW_i^T \), \( V_i = SW_i^V \) in which \( W_i \) is a parameter of linear projections and \( Q \), \( K \), \( V \) stand for learnable queries, keys, and values, respectively. A multi-layer Transformer encoder consists of a sequence of blocks, each of which contains the MHSAs and the layer normalization (LN) is applied before each block. To this end, the multi-layer Transformer encoder computes \( k = 1, \ldots, N \) layers forwardly using the following equation:

\[
S_k = \text{MHSAs}(\text{LN}(S_{k-1})) + S_{k-1}
\]  

(4)

**Unilateral cross-attention.** The proposed unilateral cross-attention vision Transformer is inspired by CrossViT [5]. However, unlike the multi-scale feature cross-attention that is symmetrical...
for two branches with small patches and large patches, the unilateral cross-attention fuses the CLS token $X_{d_b}$ of the face mask branch and feature patches from the unmasked face branch using the following equations:

$$
S_f = [f(X_{d_b})]|_{\text{patch}}^m
$$

$$
q = S_f W_q, k = S_f W_k, v = S_f W_v,
$$

$$
A = \text{softmax}(qk^T / \sqrt{C/N_k}), CA(S_f) = AV.
$$

where $W_q$, $W_k$, $W_v$ are learnable parameters, $C$ and $N_k$ are the embedding dimension and number of heads. $CA$ stands for cross attention, and $A$ is the attention map in cross-attention. Finally, the $X_{d_b}$ processed by cross attention together with feature patches from the unmasked face branch are concatenated to form the final output of each encoder layer, where $f()$ and $g()$ are the projection and back-projection function for dimension alignment, respectively. In this way, the processed CLS token serves as a query token not only to the feature patches from the unmasked face branch, but also from the face mask branch. Finally, we apply multi layer perceptron (MLP) to process the last CLS token to calculate the expression probability scores for each type of expression.

4. Experimental studies

4.1. Datasets

**M-CelebAMask-HQ.** The CelebAMask-HD [25] is a face image dataset for face parsing with 30,000 high-resolution face images, all images of which are scaled to 512 × 512 and each image is manually-annotated with 19 classes of facial components or accessories such as eyes, eyebrows, nose, lip, hair, neck, eyeglasses and clothes. The CelebAMask-HQ is processed to randomly add one of eight types of face mask on each image with the AWFM method and the images that can not be processed are discarded. Further, the label of the mask is also added and the labels of the mouth and lips are discarded. Finally, an M-CelebAMask-HQ with 28,000 face images are obtained.

**M-ibugmask.** The ibugmask [28] is a face image dataset for face parsing tasks with 21,866 images as the training set and 1000 images as the testing set. Compared with the CelebAMask-HQ, the ibugmask dataset contains more face images with large variations in the head poses. Likewise, the same process that was conducted on the CelebAMask-HQ was also conducted on the ibugmask. Finally, we also obtained a face-mask-aware face parsing dataset with 21,234 images called M-ibugmask.

**M-LFW-FER.** The M-LFW-FER [49] is a face image dataset for the face-mask-aware FER task, which we previously reported. From the dataset, 10,487 out of 13,000 samples were selected and then processed with the AWFM method so that all the face images were wearing a face mask. The images were manually annotated according to three types of facial expressions: 1,094 images for negative expressions, 5413 images for positive expressions, and 3980 images for neutral expressions.

**M-KDDI-FER and M-FER-T.** The M-KDDI-FER and the M-FER-T [49] are a face-mask-aware three facial expressions recognition training dataset and a laboratory test dataset. These datasets were used in our previous study but are not publicly available. The M-KDDI-FER contains 10,385 face images from 12 Asian subjects (5 females, 7 males) with three types of facial expressions: 3,160 images for negative expressions, 3410 images for positive expressions, and 3815 images for neutral expressions. The M-FER-T contains 562 facial images with real face masks crawled from the Internet, with 213 images for neutral expressions, 162 images for positive expressions, and 187 images for negative expressions.

**M-FER-2013.** The FER-2013 dataset [12] contains 35,887 facial images (including test set) in grayscale sized 48 × 48 and seven types of expressions: angry, disgust, fear, happy, neutral, sad, and surprise. Because it is hard to apply the AWFM method to FER-2013 images for sake of low image resolution, thus we scale up the image and then fit the face mask to the images to form a new dataset M-FER-2013 with 35,609 images: 4,908 images for angry expressions, 546 images for disgust expressions, 5074 images for fear expressions, 8951 images for happy expressions, 6164 images for neutral expressions, 5991 images for sad expressions, and 3975 images for surprise expressions.

**M-CK+.** The CK+ dataset [33] contains 593 video sequences from 123 subjects. The sequences length varies from 10 to 60 frames. Among these videos, 327 sequences from 118 subjects are labeled with seven basic expression labels and contempt: anger, disgust, fear, happy, neutral, sad, surprise, and contempt based on the facial action coding system (FACS). The dataset also comprises a total of 5876 labelled images, mostly grayscale and 640 × 490 pixels, ranging from neutral to peak expressions. In total, we collected 1587 images as a test dataset with seven emotions: 151 images for angry expressions, 185 images for disgust expressions, 104 images for fear expressions, 156 images for happy expressions, 298 images for neutral expressions, 101 images for sad expressions, 238 images for surprise expressions.

4.2. Implementation details

We use Pytorch as the deep learning framework and conduct all of our experiments on four Tesla M40 GPU cards.

**FFP model training.** We used the EAHNet model [34] for face parsing. First, all the images from the CelebAMask-HQ dataset are resized to 224 × 224 pixels, and a pre-trained EAHNet model is obtained as EAHNet-base. Then, the CelebAMask-HQ is processed by AWFM with eight types of face masks to form a new dataset M-CelebAMask-HQ. Accordingly, the label file of the mask is added, and the label files of covered components, such as the mouth, and lips, are removed. Further, we fine-tune the pre-trained EAHNet model on the M-CelebAMask-HQ to obtain a basic face-mask-aware face parsing model, which we name M-EAHNet. To take into account the head poses, we later resize all the images of the ibugmask dataset to a standardized size of 224 × 224 and make the format of the label files the same as that of the CelebAMask-HQ, and label manipulation is repeated to form the M-ibugmask dataset.

Based on M-EAHNet, the EAHNet mode is fine-tuned again on the M-ibugmask dataset to obtain another model M-EAHNet(fine-tune). To make the face-mask-aware face parsing model more compatible with grayscale images, M-CelebAMask-HQ and M-ibugmask are also transformed to grayscale images and fine-tuning is also conducted using the same procedure to obtain EAHNet-base-gray, M-EAHNet-gray and M-EAHNet(gray-fine-tune), respectively.

We then train the model for 300 epochs on the M-CelebAMask-HQ, and 200 epochs on the M-ibugmask. The batch size is set to 16 for all FFP models. The datasets of M-CelebAMask-HQ and M-ibugmask are divided into training data and validation data in a ratio of 7:3. To avoid over-fitting, common data augmentations are used: random horizontal flip, random color jittering, random crop image, etc. The learning rate starts at 1e−3, following a momentum of 0.9 and a weight decay of 1e−5. Finally, the 2D cross-entropy is defined as the loss function. The code of EAHNet in Github is referenced.1

**VTC-FER model training.** First, the MTCCN [53] is utilized to detect face images from the image to obtain the face cropped version. Then, the face image is resized to 224 × 224 and fitted with a face mask using the AWFM method. Subsequently, the tensor of an image is manipulated with an element-wise product by the mask

---

1 https://github.com/tracelessLe/FaceParsing.PyTorch.
map of the face mask and the mask map of unmasked facial skins to form a 2D tensor for two branches. Afterwards, the backbone of ResNet50 [13], which is pre-trained on ImageNet [8], is employed to extract features from the 2D tensor for each branch.

We train the VTC-FER model for 300 epochs with a warmup of 50 epochs, and the batch size is set to 32. The datasets of M-LFW-FER, M-KDDI-FER, and M-Ck+ are divided into training data and validation data in a ratio of 7:3. To avoid over-fitting, common data augmentations are used: random horizontal flip, random color jittering, random crop image, etc. For the multi-layer Transformer encoder (MTE), the number of layers and heads are set to 4 and 8 respectively. The input features of MTE are reshaped into patches with a size of (12, 16), and its hidden dimension is set to 256. Finally, the hidden dimension of the MLP header is set to 1000. The learning rate starts at $5e^{-3}$, following a warmup of 500 steps and a cosine learning rate decay. The AdamW [23] is employed to optimize the whole network, and the standard cross-entropy loss is utilized as the loss function.

4.3. Results comparison

**Evaluation of FFP model.** The evaluation of FFP model can be roughly divided into two parts: First is the overall evaluation and the other one is the class performance evaluation, as is shown in Table 1. For the overall evaluation, we mainly consider the overall accuracy, mean IoU, and overall F1-score. For the class performance evaluation, since we are only concerned about the separation accuracy of the face mask and unmasked facial skin, only the mask IoU and the skin IoU are shown in the table. Considering the computational cost, the amount of floating-point operations (FLOPs) and parameters (#Params) of the M-EHANet(fine-tune) and the M-EHANet-gray(fine-tune) are also listed in the table for reference.

For the modes dealing with RGB images, the overall accuracy of M-EHANet and M-EHANet(fine-tune) has been improved compared with the pre-trained model EHANet-base. This improvement could be the result of the elimination of lips and mouth labels and the highly accurate prediction of face mask. Mean IoU and overall F1-score of M-EHANet and M-EHANet(fine-tune) have decreased slightly compared with EHANet-base but maintained almost the same value, which means taking face mask into account does not decrease the overall performance of the model. For the mask IoU of M-EHANet and M-EHANet(fine-tune), we can find that they all reached about 0.98 to 0.99, which means that prediction of the mask region is highly accurate at the pixel level. The skin IoU of M-EHANet and M-EHANet(fine-tune), which also have to take the face mask into account, reached about 0.88 to 0.89, compared with 0.92 for EHANet-base. For the modes dealing with grayscale images, we also follow the trend of the models dealing with RGB images. However, the overall performance and class performance of EHANet-base-gray, M-EHANet-gray, and M-EHANet-gray(fine-tune), have decreased compared with that of EHANet-base, M-EHANet, and M-EHANet(fine-tune), accordingly.

**Evaluation of VTC-FER model on three expressions.** As is shown in Table 2, VTC-FER is evaluated on three expressions. The table consists of three blocks: 1) Well known pre-trained model backbones are listed in the first block for comparison, such as VGG19 [29], MobileNetV2 [42], ResNet50, and vision Transformer (ViT). 2) Existing occlusion-aware FER approaches are listed in the second block for comparison, such as RAN, ACNN, and OADN introduced in section “2.2.” 3) Our previous method and newly proposed method are listed in the third block for comparison.

All the models are trained on the M-LFW-FER dataset or the M-KDDI-FER dataset and then tested separately on the M-KDDI-T dataset. Compared with pre-trained VGG19, MobileNetV2, ResNet50, and vanilla vision Transformer, our proposed method demonstrates a major improvement in facial expression recognition performance. Meanwhile, it is apparent that the occlusion-aware FER approaches performed better than the normal classification models in terms of recognition accuracy on both the M-LFW-FER and M-KDDI-FER datasets, as well on the test dataset M-KDDI-T. Among the existing occlusion-aware FER approaches, the overall detection accuracy of OADN was better than RAN and ACNN, for it proposed a region partition branch to consider the large occluded region. Compared with OADN, RAN, and ACNN, our previously proposed method achieved the best performance in terms of the accuracy evaluation on both datasets, as well as on the test datasets. And the newly proposed method outperformed the previously proposed method in terms of the evaluation in both the training and testing phases. This is mainly because the proposed FFP model first increased the separation accuracy of the unmasked facial region, and then the proposed VTC-FER model balanced the attention on the unmasked facial region and masked facial region to obtain an optimized face-mask-aware FER model. The confusion matrix images of the proposed method on the M-LFW-FER, the M-KDDI-FER, and the M-LFW-T datasets are shown in Fig. 3.

Considering the computational cost comparison of each method listed in Table 2, it could be found that the proposed method demonstrated its superiority in terms of both FLOPs and #Params, compared with most existing methods. In particular, the FLOPs decrease exponentially compared with our previous study (13.7G vs 5.3G). However, the price we paid is that the number of parameters for the proposed model is larger than our previously proposed method (23.5M vs. 19.3M).

**Evaluation of VTC-FER model on seven expressions.** For accurate evaluation of face-mask-aware FER on datasets with seven expressions, we mainly consider two processed datasets: the M-FER-2013 and the M-Ck+, as is shown in Table 3. The proposed model is trained on the M-FER-2013 and tested on the test set of M-FER-2013 and the M-Ck+, separately. Likewise, the table also consists of three blocks like the table of accuracy evaluation on
three expressions. In the third block, CroppedFace stands for the approach in [4], which is also the state-of-the-art (SOTA) on seven expressions to the best of our knowledge. Previous stands for our previous method, while proposed means the method proposed in this paper. Although our previous method was not evaluated on datasets with seven expressions, it still outperforms the CroppedFace method on both the M-FER-2013 and M-CK+ datasets. And our newly proposed method outperforms both the existing SOTA and our previously proposed method on both datasets. It was shown that the existing SOTA, which uses the eye region for recognition but discarded some important parts, decreases the detection accuracy.

The confusion matrix images of the proposed method on the M-FER-2013 and CK+ datasets are shown in Fig. 3. In particular, the recognition accuracy of SOTA on the M-FER-2013 dataset with angry, disgust, fear, happy, neutral, sad, and surprise are 0.46, 0, 0.59, 0.58, 0.49, 0.41, 0.68. Our proposal obtained the following results: 0.66, 0.34, 0.66, 0.72, 0.67, 0.59, 0.70, outperforming the SOTA method in terms of detection accuracy of all the expressions. Likewise, The experimental results of the proposed method tested on M-CK+ also outperformed the SOTA method in terms of detection accuracy of all the expressions.

Considering the computational cost comparison of each method in Table 3, it could be found that the proposed method demonstrated its superiority in terms of both FLOPs and #Params, compared with most existing methods. In particular, the proposed method did not demonstrate its superiority compared with the CroppedFace in terms of both FLOPs and #Params. However, the detection accuracy of the proposed method is improved a lot on both datasets of M-FER-2013 (0.6653 vs. 0.5428) and M-CK+ (0.6108 vs. 0.5079).

**Evaluation Summary.** To study the performance of the proposed model in-depth, we first take the backbone of VGG19, MobileNetV2, ResNet50, and ViT for comparison. In particular, the
The Transformer ViT model is trained without face parsing and the cross-attention mechanism showed that there was a big gap in performance compared to our proposed methods. Other than these classical models, experiments with models devised for occlusion-aware facial expression recognition, such as RAN, ACNN, and OADN, were also conducted for comparison. Lastly, the performance of two existing face-mask-aware FER methods, among which one is our previous work, are also evaluated. According to the evaluation results, the conclusion can be drawn that our proposed method outperforms all existing methods in terms of face-mask-aware facial expression detection accuracy. Some correctly and incorrectly predicted facial expression samples by our proposal are shown in the Fig. 5 for reference.

4.4. Ablation studies & insights

As is described in this paper, our proposal mainly consists of two important parts: 1. face-mask-aware face parsing. 2. vision Transformer with a cross-attention mechanism. To evaluate the effect of the proposal, each part is tested and compared with our previous proposal. The evaluation results are shown in Table 4.

The impact of face-mask aware face parsing. As is shown in Table 4, BC stands for the binary classifier that is used to locate the upper face as presented in our previous study, while FFP stands for the face-mask-aware face parsing in this paper. They are all conducted in the first stage to locate the non-occluded facial region. First, we implement the binary classifier AD in our previous work in the second stage. Then we shift the BC to FFP in the first stage and find that detection accuracy increases by 0.87% from 87.92% to 88.79% on M-LFW-FER, and the accuracy increases by 1.25% from 61.79% to 63.04% on M-FER-2013. Then VTC-FER is implemented in the second stage. We shift the BC to FFP and find that the detection accuracy increases by 0.85% from 89.46% to 90.31% on M-LFW-FER, and the accuracy increases by 1.56% from 64.97% to 66.53% on M-FER-2013. From the evaluation, the conclusion can be drawn that face-mask-aware face parsing contributes to improving the performance of FER. As is discussed in Subsection 3.1, our hypothesis that locating the unmasked facial regions precisely will increase
the performance of FER is proved by the ablation study, compared with the binary classifier that only excludes the region of the face mask.

The impact of ViT with cross-attention mechanism. To study the impact of ViT-FER in the second stage, we use BC in the first stage and shift AD to ViT-FER in the second stage. By doing this, the detection accuracy increases by 1.54% from 87.92% to 89.46% on M-LFW-FER, and the accuracy increases by 3.18% from 61.79% to 64.97% on M-FER-2013. Then we use FFP in the first stage and shift AD to VTC-FER in the second stage. By doing this, the detection accuracy increases by 1.52% from 88.79% to 90.31% on the M-LFW-FER, and also increases by 3.49% from 63.04% to 66.53% on the M-FER-2013. From the evaluation, the conclusion can be drawn that a vision Transformer with a cross-attention mechanism contributes to improving the performance of FER. It can also be seen that the ViT-FER part makes a greater contribution to improving accuracy than the FFP part. As discussed in Subsection 3.2, the ablation study also proved that the vision Transformer with a cross-attention mechanism could adaptively re-weight attention on the masked and unmasked facial regions to achieve the optimal performance of the proposed model, rather than empirically assign a specific percentage to each region.

5. Limitations

Generality. Although our proposal out-performs our previous work as well as occlusion-aware FER methods in terms of face-mask-aware FER, the ViT-FER is designed to deal with the face mask problem in FER in particular. Our proposal is therefore evaluated on the real-world occluded dataset FED-RO [27] to evaluate its limitation, as shown in Table 5. The FED-RO is a publicly available test dataset containing 400 occluded images with seven expressions. The proposed method along with our earlier method is trained on M-FER-2013 and tested on the FED-RO dataset. Our proposals did not out-perform general occlusion-aware FER methods. Thus, it could be said that our proposal achieves the best performances in terms of face-mask-aware FER by sacrificing the generality to some extent.

We discovered that facial image inpainting methods are able to restore face images even when parts of the faces are masked by irremovable occlusions, as shown in several works [6,17,24,37,52]. This remains as a potential solution for general occlusion-aware FER, which could be a topic for consideration in a future study.

Practicality. Our earlier proposals on three expressions have already been applied in practice [51]. However, in the actual use, while we found that the proposal was able to demonstrate good performance for macro-expressions, it was lacking in sensitivity for micro-expressions. This remains a problem that needs to be solved in a future study. Furthermore, although our proposal shows a SOTA evaluation results compared with several existing works on M-FER-2013 and M-CK+, we found that a big gap still exists in terms of practicality when tested in the real applications. This could be because some expressions are highly linked to the region around the mouth and nose that are fully covered by a face mask, a problem we referred to in the introduction. There are also several studies about the effect of facial masks on the perception of facial emotion [40,62]. Thus, the practicality of face-mask-aware FER on more than three expressions, such as six and seven expressions, should be addressed in a future study. Other than the facial expression, behavioral modalities such as body gestures and speech, or physiological modalities such as electroencephalography (EEG) and galvanic skin response (GSR), could be better options to understand detailed human emotions in the case of masked face.

6. Conclusion

In this paper, we proposed an advanced face parsing and vision Transformer along with a cross-attention-based method to improve the detection accuracy of face-mask-aware FER. The proposed method consists primarily of two parts: the face-mask-aware face parsing model and the vision Transformer along with a cross-attention mechanism. The proposed method outperforms existing state-of-the-art face-mask-aware FER methods, as well as other occlusion-aware FER methods, on both datasets with three kinds of emotions and datasets with seven kinds of emotions, whether in the wild or in the laboratory.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Parts of the datasets used in the paper are already publicly available, other datasets and codes are confidential.

References

[1] M.A.H. Akhand, S. Roy, N. Siddique, M.A.S. Kamal, T. Shimamura, Facial emotion recognition using transfer learning in the deep CNN. Electronics (Basel) 10 (9) (2021), doi: 10.3390/electronics10091036.
[2] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.-P. Morency, OpenFace 2.0: Facial behavior analysis toolkit, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG) 2018, pp. 59–66, doi: 10.1109/FG.2018.00019.
[3] C.C. Carbon, Wearing face masks strongly confounds counterparts in reading emotions. Front. Psychol. 11 (2020) 2526, doi: 10.3389/fpsyg.2020.939433.
[4] G. Castellano, B. De Carolis, N. Macchiariulo, Automatic emotion recognition from facial expressions when wearing a mask, in: CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter, in: ACM CHItaly 2021, Association for Computing Machinery (ACM), 2021, pp. 1–5, doi: 10.1145/3464385.3464740.
[5] C.-F. Chen, Q. Fan, R. Panda, Crossvit: cross-attention multi-scale vision transformer for image classification, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 347–356.
[6] Y.-A. Chen, W.-C. Chen, C.-P. Wei, Y.-C.F. Wang, Occlusion-aware face inpainting via generative adversarial networks, in: 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 1202–1206, doi: 10.1109/ICIP.2017.8294672.
[7] X. Cheng, J. Lu, B. Yuan, J. Zhou, Face segmentor-enhanced deep feature learning for face recognition, IEEE Trans. Biometric. Behav. Identity Sci. 1 (4) (2019) 223–237, doi: 10.1109/TBIB.2019.2936624.
[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagnet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
[9] H. Ding, P. Zhou, 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, doi: 10.1109/IJCB48548.2020.9104923.
[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021, https://openreview.net/forum?id=YicbflNtQg.
[11] M.-I. Georgescu, R.T. Ionescu, Recognizing facial expressions of occluded faces using convolutional neural networks, ICONIP, 2019.
