Transfer-Learning Enabled Micro-Expression Recognition Using Dense Connections and Mixed Attention

Chenquan Gan[a,b], Junhao Xiao[b], Qingyi Zhu[a], Deepak Kumar Jain[c], Vitomir Štruc[d,*]

[a] *School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
[b] *School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
[c] *Institute of Automation, Chongqing University of Posts and Telecommunications Chongqing 400065, China*
[d] *Faculty of Electrical Engineering, University of Ljubljana, Trzaska cesta 25, SI-1000 Ljubljana*

Abstract

Micro-expression recognition (MER) is a challenging computer vision problem, where the limited amount of available training data and insufficient intensity of the facial expressions are among the main issues adversely affecting the performance of existing recognition models. To address these challenges, this paper explores a transfer–learning enabled MER model using a densely connected feature extraction module with mixed attention. Unlike previous works that utilize transfer learning to facilitate MER and extract local facial-expression information, our model relies on pretraining with three diverse macro-expression datasets and, as a result, can: (*i*) overcome the problem of insufficient sample size and limited training data availability, (*ii*) leverage (related) domain-specific information from multiple datasets with diverse characteristics, and (*iii*) improve the model adaptability to complex scenes. Furthermore, to enhance the intensity of the micro-expressions and improve the discriminability of the extracted features, the Euler video magnification (EVM) method is adopted in the preprocessing stage and then used jointly with a densely connected feature extraction module and a mixed attention mechanism to derive expressive feature representations for the classification procedure. The proposed feature extraction mechanism not only guarantees the integrity of the extracted features but also efficiently captures local texture cues by aggregating the most salient information from the generated feature maps, which is key for the MER task. The experimental results on multiple datasets demonstrate the robustness and effectiveness of our model compared to the state-of-the-art.

Keywords: Micro-expression recognition, transfer learning, dense connections, mixed attention, Euler video magnification

1. Introduction

Facial expressions are the primary means of expressing emotions in day–to–day interactions [1, 2, 3]. *Micro-expressions* represent a special type of facial expression that corresponds to involuntary facial move-

[*] Corresponding author

Email addresses: gcq2010cqu@163.com (Chenquan Gan), 402332840@qq.com (Junhao Xiao), zhuqy@cqupt.edu.cn (Qingyi Zhu), deepak@cqupt.edu.cn (Deepak Kumar Jain), vitomir.struc@fe.uni-lj.si (Vitomir Štruc)

ments triggered by emotional stimuli [4]. Micro-expressions not only reflect the hidden emotions of human beings in extreme situations, but can also be used to validate their authenticity [5]. The development of automatic micro-expression recognition (MER) techniques has led to the successful deployment of MER technology in criminal investigations, job interviews, and clinical medicine among others [6]. Unfortunately, the extremely short duration (from 1/25 to 1/3 seconds) of micro expressions, their weak intensity, and sparsity across time all contribute to the difficulty of the research on this topic [7].

Initial studies on MER focused predominantly on the analysis of complete video clips [8, 9], resulting in time-consuming analyses that had to deal with data redundancy and complex models capable of extracting micro-expression information from sequences of frames. Following the insights from Ekman [10], later techniques (e.g., [11]) shifted attention to the analysis of so-called apex frames, i.e., video frames corresponding to the peak intensity of the facial expressions, which are now generally considered to be better suited for an automated analysis of the micro-expressions. These techniques not only address data redundancy in an explicit manner, but also lead to computationally simpler recognition models. Although the handcrafted features utilized with early apex-frame based methods performed reasonably well, the overall performance still warranted additional research efforts.

With the excellent performance of deep learning and convolutional neural networks (CNNs) in face recognition and other face-related vision tasks [12, 13, 14], deep-learning-based methods also received widespread attention for the micro-expression recognition problem [15]. While pioneering (deep learning) work in this area produced only modest performance improvements compared to prior techniques due to the limited amount of training data available, subsequent works reported better results by exploring different strategies. The work in [16, 17], for example, aimed to mitigate the impact of the small number of samples available for training by using shallow network/model architectures with smaller numbers of parameters that could be estimated reliably from limited training data. However, the features extracted from such models were shown to be inferior to the features extracted from deeper models. In the pursuit of optimizing MER model performance, several innovative works have been applied to MER in recent years. These include, for instance, transformer-driven MER models [18], MER approaches based on dual-stream (local and global) attention [19], and MER frameworks leveraging local facial behaviors learning from enhanced expression flow [20], among others. The work in [21, 22, 23] used transfer learning to eliminate the impact of insufficient training data and achieved good results. However, unlike the application of the transfer learning in other areas [24, 25, 26], source domain datasets [21] with an insufficient overlap in terms of data characteristics of the MER task as well as features with limited expressive power [22, 23] still affected the final performance. In [27], the Euler video magnification (EVM) algorithm was adopted to enhance the intensity of facial expressions and, consequently, to improve the feature strength by amplifying the motion in video. A spatial attention mechanism was also used to contribute toward the acquisition of important local texture features in [28]. While the effectiveness of motion amplification and attention mechanisms on MER was demonstrated in these works, important (local) texture information, key for recognizing micro-expressions, was still lost during the feature extraction process.

From the above discussion, it follows that answers to the following two key questions are critical for further improving MER performance:

1) How can transfer learning be utilized better to accomplish MER?

2) How can the feature extraction model better capture local texture information while minimizing the loss of important MER cues in the computed feature representations?

To address the first question, existing transfer-learning methods either used: (*i*) standard ImageNet pretraining, e.g., [21], which often led to sub-optimal MER performance due to the mismatched between the source and target domain, or (*ii*) a single macro-expression dataset to initialize the model parameters for MER, e.g., [22, 23], which adversely affected the adaptability of the learned MER model and its applicability to micro-expression data captured in different environments and settings. With the model proposed in this paper, we improve on the outlined solutions, by merging three diverse macro-expression datasets into a mixed dataset for model pretraining. This strategy not only better captures the input data variability but also improves the multi-scene adaptability of the final model. Furthermore, we use the EVM algorithm to improve the intensity of micro-expressions, which further reduces the appearance gap with the macro-expression source datasets.

To extract discriminative local features, prior methods used simple spatial attention mechanisms and feature extraction structures that typically resulted in a considerable loss of discriminative information [28]. To address these issues, we propose a novel feature extraction approach in this work that combines densely connected structures with a novel mixed (*channel-spatial*) attention mechanism, ensuring the integrity of the output features and the ability of the model to focus computational resources on discriminative cues from the input data.

In summary, we make the following main contributions to this paper:

- We propose a novel strategy for MER model pretraining that exploits multiple (diverse) macro-expression datasets and ensures that a reliable, adaptive, and competitive MER model can be learned using limited amounts of (micro-expression) training data. Additionally, we use Euler Video Magnification (EVM) to further improve the correspondence between the source (i.e., macro expression) and target (i.e., micro-expression) domains.

- We introduce a novel feature extraction approach built around a densely connected feature extraction module and a mixed (channel-spatial) attention mechanism that is capable of extracting highly discriminative features for micro-expression recognition.

- We show the benefit of using the proposed pretraining strategy and feature extraction module for the MER task in comparative experiments with state-of-the-art techniques from the literature and report highly competitive performance on multiple benchmarks.

The rest of the paper is organized as follows: In Section 2, works in the literature that are close to our proposed approach are summarized; In Section 3, our proposed approach is presented; A comprehensive experimental evaluation is reported in Section 4; Conclusions and future research directions are discussed in Section 5.

2. Related Work

Since the pioneering work of Pfister *et al.* [29], research on micro-expression recognition (MER) has seen significant attention from the computer-vision and affective-computing communities. Early studies in this area relied heavily on handcrafted features to solve the MER task. Li *et al.* [30], for example, proposed to approach the problem by combining local binary pattern (LBP-TOP) from three orthogonal planes with traditional off-the-shelf classifiers. Inspired by the success of LBP-TOP, other researchers proposed extensions of the original technique, and improved the recognition accuracy by reducing the redundancy of the LBP-TOP operator [31]. To improve the accuracy of MER, new image descriptors, such as SCCLQP [9] and HIGO-TOP [32], were also proposed based on the LBP-TOP features. Concurrently, novel features, such as MDMO [33], were introduced and observed to ensure competitive performance. Most of these methods were applied to video sequences and, as a consequence, were also computationally expensive. The work in [10] analyzed the process of micro-expression recognition from a psychological perspective and found that the information conveyed by micro-expressions at their peak intensity is highly reflective of the current emotional state. Liong *et al.* [34] verified this assertion through comprehensive experiments and proposed using apex frames instead of video sequences as the basis for automated MER. Unfortunately, the performance of this early apex-frame based model did not surpass the best-performing video-based techniques. Nonetheless, the introduction of the concept of apex frames represented a major milestone for MER research.

Given the advances in deep learning and its impressive results in various problem domains [35, 36, 37, 38, 39], researchers started looking increasingly at deep-learning solutions to improve the accuracy of micro-expression recognition. Platel *et al.* [15], for instance, applied a deep convolutional model to MER. With this approach, a deep convolutional neural model used earlier for face recognition was adapted for micro-expression recognition using a single target dataset. However, due to the lack of available training data in the single target dataset, the learned model failed to meet expectations. Consequently, the accuracy of this model on the CASME II dataset was lower than the accuracy of methods relying on handcrafted features. Peng *et al.* [40] proposed the dual-branch CNN network for MER and provided an important experimental basis for the future study of shallow networks trained with limited amounts of data. Quang *et al.* [41] introduced a capsule network to solve the small sample size problem. Liong *et al.* [16] designed a shallow three-stream three-dimensional network (STSTNet) in combination with optical flow characteristics. By processing the onset frame and apex frame of the input samples, three characteristics of optical strain, horizontal optical flow, and vertical optical flow were obtained and finally classified. Gan *et al.* [42] proposed to use optical flow features from the apex frame network (OFF-ApexNet) for MER. OFF-ApexNet first extracted optical flow features corresponding to micro-expressions in the apex frames, and then performed feature extraction through a CNN model. Later, Xia *et al.* [17] proposed using a combination of low-resolution inputs and recursive convolution networks (RCNs) to emulate the training characteristics of a shallow network with a comparably deeper model. Although a series of shallow networks were proposed in the works discussed above, which resulted in notable performance improvements, MER models still exhibit weaker performance than comparable models used in other visual recognition tasks.

In order to further reduce the impact of the small sample sizes available for training, Peng *et al.* [21] used transfer learning to provide a better starting point for model fine-tuning. However, due to the ImageNet pretraining, the correspondence between the feature distributions of the source and the target domain samples was insufficient, adversely affecting the final results. Later, Ben *et al.* [22, 23] proposed two transfer-learning methods based on LBP features. Although good results were reported, the main problem with these approaches was that they over-considered local information, while ignoring the overall appearance information. In this paper, we aim at (*i*) expanding the amount of available training data, (*ii*) increasing the relevance between the source and target domains during transfer learning, and (*iii*) improving the adaptability of the model to multiple scenarios and settings, and propose to utilize three macro-expression datasets from different environments for pretraining, and then to fine-tune the pretraining model for the MER task. Additionally, inspired by the work in [27], we also propose to integrate the EVM algorithm into the transfer-learning procedure to further reduce the mismatch between macro- and micro-expression data.

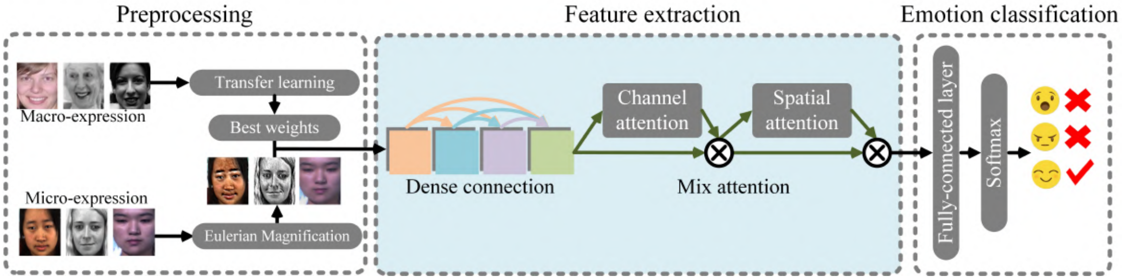


Figure 1: Overview of the proposed micro-expression recognition (MER) method. The method relies on a powerful transfer learning strategy using multiple macro-expression datasets and Euler Video Magnification (EVM) and a novel feature extraction technique that uses a densely connected feature extraction module and mixed attention to derive a rich and descriptive set of image features for recognition. The figure is best viewed in color.

For feature extraction, Zhou *et al.* [28] used a model that combined a ResNet with spatial attention. While promising results were reported, the extracted features still lacked expressiveness and discriminability. Furthermore, because features based on optical flow are highly suitable for encoding motion information, such features were still favored by the majority of researchers. Some works chose to extract optical flow features from the entire video segment [15, 32, 40, 43], while others used onset frames, apex frames, and offset frames for calculating optical-flow based representations [44, 45]. Even though competitive results were achieved, the lack of local texture information and the high computational complexity are among the main shortcomings of such techniques. Conversely, in this paper, we design a feature extraction model that combines a densely connected structure and a mixed attention mechanism to alleviate most of the limitations discussed above. The model ensures that highly descriptive features are extracted from the input data through the densely connected processing structure, and the low number of parameters allows for the design of a model with sufficient depth. At the same time, the addition of the mixed attention mechanism also provides the model with a significant ability to capture local texture information. As we show in the experimental section, our design leads to highly competitive recognition performance when compared to state-of-the-art methods from the literature.

3. The Proposed Method

This section first provides an overview of our method and then introduces the preprocessing part and feature extraction module of the proposed approach in detail. Finally, the emotion classification module of the model is described.

3.1. Method Overview

As illustrated in Figure. 1, the proposed method consists of three main stages aimed at: (*i*) preprocessing and transfer learning, (*ii*) feature extraction, and (*iii*) micro-expression (emotion) recognition. In the *first stage*, the input data is preprocessed and the transfer-learning approach with the Euler Video Magnification (EVM) algorithm is used. Next, pretrained weights (learned from macro-expression datasets) are utilized to initialize the MER model, which is then fine-tuned using the available micro-expression data. In the *second stage*, the EVM-enhanced apex frames (corresponding to a micro-expression) are fed as input to the feature extraction model. The feature extraction model itself is divided into two parts: (*i*) the first is a densely connected feature extraction module, and (*ii*) the second is a mixed attention module. The feature extraction model utilizes a densely connected module as its backbone to extract facial representations pertinent to the MER task. The design of this model facilitates feature reuse and allows our model to learn highly descriptive features for MER without losing important information [46]. We deploy the mixed attention module into the skip pathways of the densely connected module to focus computational resources on emotional information, thereby improving the performance of the feature extractor. In the mixed attention module, the channels of the extracted feature maps are first reweighted according to their importance. Next, the channel-weighted feature maps are fed to the spatial attention module, which again weights the entries with respect to their importance, however, this time, the weighting is performed across the spatial dimension. In the *last stage*, the features are passed as input to the classification module that makes the final decision regarding the micro-expression class through a final processing step implemented by a fully-connected layer and the softmax classification function.

3.2. Preprocessing

In the process of MER, there are two major challenges: (*i*) The first is the small number of samples available in existing micro-expression datasets, and (*ii*) the second is that micro-expressions have a lower range of motion compared to macro-expressions.

In order to address challenge (*i*) and (*ii*), a transfer learning method is adopted in the preprocessing part of our approach. Here, three macro-expression datasets, i.e., CK+ [47], RAF-DB [48], FER-2013 [49], are used to conduct pretraining and find a good initialization for the parameters of the recognition model. These pretrained weights are then used for initialization before the training (fine-tuning) on the micro-expression datasets. Because both macro- and micro-expressions are produced by facial deformations caused by the synergistic action of 42 facial muscles, the characteristics of the appearance changes of macro- and micro-expressions are very similar. Consequently, pre-training on the macro-expression datasets can allows the feature extractor to learn (meaningful and informative) prior data representations, that eventually lead to improved performance on the targeted micro-expression datasets after finetuning.

The approach leverages apex frames as input instead of entire video sequences, with the goal of improving training efficiency by reducing the temporal complexity of the input data for the feature extractor. Additionally, the EVM algorithm, which is instrumental in tackling challenge (*ii*), significantly strengthens the expresivnes of the micro-expression samples. This enhancement reduces the domain divergence between the source domain (macro-expressions) and the target domain (micro-expressions), which is essential for the efficacy of transfer learning. Specifically, with the proposed approach, multi-frame images in the micro-expression datasets are first sampled from the input video sequences. This is done to cater to the characteristics of EVM, which requires multi-frame data as input. Next, the EVM algorithm is used to process all of the sampled multi-frame images. To avoid amplifying environmental noise, a target frequency-band that ensures that only the desired motion is magnified needs to be chosen. Since the ideal frequency range for micro-expression movements is usually from 0.1 Hz to 0.4 Hz [50], we set this frequency range as the amplification range of the EVM algorithm.

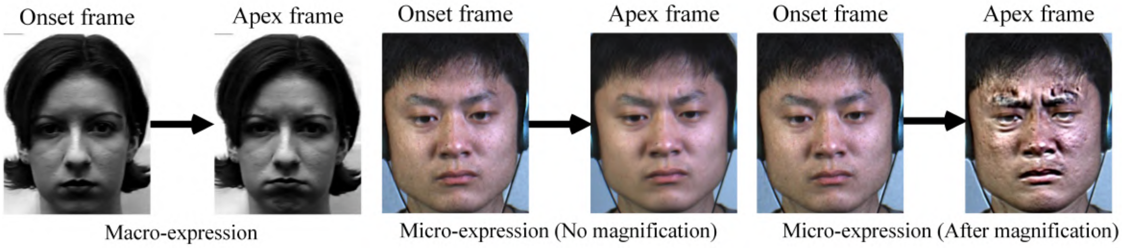


Figure 2: Comparison of macro-expressions and micro-expressions (before and after) Euler Video Magnification. The use of the magnification technique serves a dual purpose in our work: (*i*) it improves the intensity of micro-expressions making them easier to recognize, and (*ii*) it reduces the appearance difference between macro- and micro-expressions, thus, facilitating efficient transfer learning.

The EVM procedure used for the presented process is formally described as follows [51]:

$$Pixel\left(x,0\right)=p\left(x\right), \tag{1}$$

$$Pixel\left(x,t\right)=p\left(x+\delta\left(t\right)\right), \tag{2}$$

$$Pi\hat{x}el\left(x,t\right)=p\left(x+\left(1+\alpha\right)\delta\left(t\right)\right), \tag{3}$$

where $Pixel\left(x,t\right)$ represents the pixel's brightness at time instance t; $Pi\hat{x}el\left(x,t\right)$ denotes the pixel's brightness at time instance t after amplification; $\delta\left(t\right)$ indicates the displacement of target motion; and α stands for the amplification factor. As illustrated in Figure. 2, compared to the original apex frame image, the intensity of the micro-expression of the apex frame processed by EVM is significantly improved. Moreover, the enhanced expression is very close in appearance to the macro-expression in the corresponding apex frame.

In the last step, a complete video is restored from the enhanced frames using the original sampling frequency. Apex frames corresponding to the micro-expression are finally selected for experimentation according to the order of the apex frames in the micro-expression datasets. The apex frames are rescaled to a size that fits the architecture of the feature extraction module. This module is described in detail in the following section.

3.3. Feature Extraction

Because the input data is processed sequentially (layer after layer) in standard convolutional networks, a certain degree of feature loss inevitably occurs due to the different types of information encoded at each of the network layers, e.g., lower network layers typically encode low–level image characteristics, whereas higher layers encode higher–level image semantics. Additionally, significant variability in appearance, attributable to varying illumination, pose changes, and other nuisance factors, often presents a challenge for a model to accurately identify and focus the location of crucial information during the feature extraction process. Therefore, it is paramount that the feature extraction process is designed in a way that mitigates these problems. Based on this insight, we propose a densely connected feature extractor with a mixed attention mechanism in the following sections.

3.3.1. Densely-Connected Module

To improve the efficiency of feature extraction and reduce the loss of features across the model layers, the proposed feature-extractor adopts a backbone that consists of stacked densely connected modules. As shown in Figure. 3, the output features of a given convolutional layer in the densely-connected module are used as the feature input of all subsequent convolutional layers. This process facilitates feature reuse down the module's layers and ensures that complementary features are learned in the additional channels of each layer. Furthermore, as the features from the lower layers are propagated to all subsequent layers, all of the encoded image information is still present at the final output layer. Thus, the adopted connection mode exhibits better feature extraction characteristics than competing architectures and improves the design of

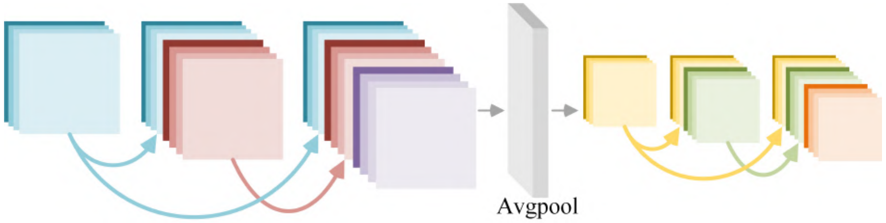


Figure 3: High–level overview of the feature extraction module implemented with a densely connected feature extractor. The proposed module allows for feature reuse in all layers and does not suffer from the loss of potentially important information down the model layers.

standard (stacked-only) feed-forward layers. In this backbone, the appearance semantics (from the lower layers) are integrated with the abstract semantics (from the higher layer) through dense skip connections, which facilitates the retention of useful information and provides comprehensive semantic information for emotion classification. Additionally, with standard models, the feature (and information) loss becomes more severe as the depth of the model increases, while the densely-connected design avoids this critical drawback.

In the densely connected module, a 5 × 5 kernel size is selected for the convolution layers, and a batch-normalization operation is utilized after each convolutional layer to normalize the distribution of features, reduce the internal covariate shift, and improve the convergence of the model during training. The result of these operations is then fed as input to the next layer. In the connections between the layers, the module exploits feature splicing, where the output features of the previous layer are directly spliced with the output of the current layer. Such a connection mechanism reduces the loss of features in the transmission process, minimizes the number of calculations needed, and reduces the number of model parameters. Additionally, after feature extraction with the given densely-connected sequence of convolutional layers (also called a *block*), the module down-samples and splices the generated feature maps using average pooling, thereby compressing and reducing the feature dimension. The output of the i^{th} convolution layer of the j^{th} dense block can be expressed as:

$$\mathrm{H}_j^i = BN\left(\sigma_1\left(Conv_{5\times 5}\left(Concat\left(\mathrm{H}_j^0, \mathrm{H}_j^1, \mathrm{H}_j^2, \ldots, \mathrm{H}_j^{i-1}\right)\right)\right)\right), \quad (4)$$

where $BN(\cdot)$ and $\sigma_1(\cdot)$ respectively present the batch-normalization layer and ReLu activation function, $Conv(\cdot)$ is a composite function that denotes all operations in each convolutional layer, $Concat(\cdot)$ denotes the concatenation operator on the channel dimension and H_j^0 is the input of the module.

3.3.2. Mixed Attention Module

Different facial expressions typically occur in different parts of the face and at varying degrees of intensity, e.g., disgust is usually accompanied by frowning and a narrow mouth, and the corresponding (active) facial areas are the eyebrows and mouth. Steering the model to allocate finite computational resources to critical facial areas not only augments the efficiency of the feature extraction process but also heightens the concentration of the model on the objective task. In addition to spatial information, the channel information of features also can assist in pinpointing the regions where the crucial facial information is. Thus, a mixed attention mechanism, focusing on spatial and channel of feature, is instrumental to precisely locating emotional regions.

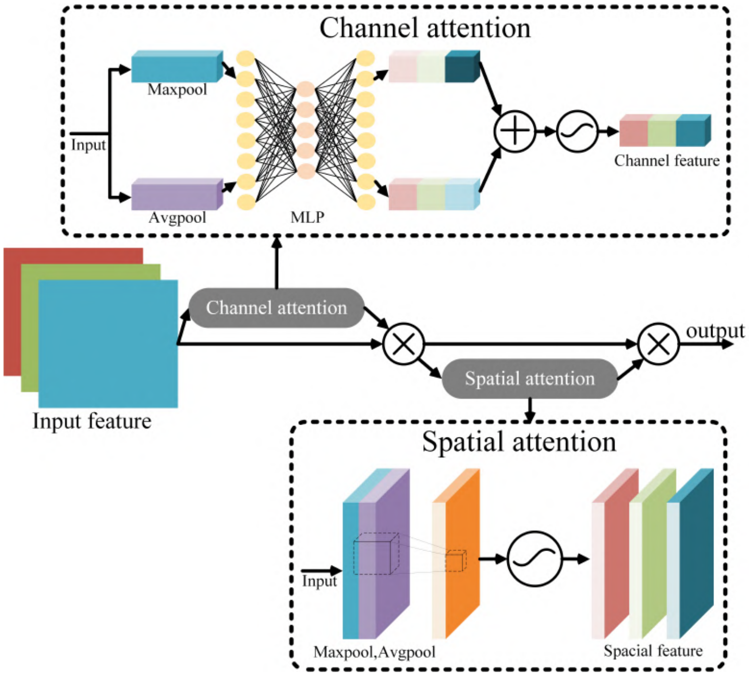


Figure 4: Overview of the mixed attention mechanism implemented in the proposed model. The mechanism implements a channel as well as a spatial attention step and ensures that the highly discriminative features are computed for the MER task by the proposed feature extraction module.

For the above reasons, we propose a mixed attention module and integrate it in the skip pathway of the densely connected module. The proposed attention module combines powerful channel-attention and spatial attention steps, as illustrated in Figure. 4. The proposed *channel-attention* mechanism is implemented with a multi-layer perceptron (MLP), which, given some input feature representation, generates a set of corresponding channel weights. The MLP is fed with two types of features, i.e., features extracted by either max-pooling or average pooling operations applied over the output of the densely connected feature extractor. This dual encoding (through the different pooling operations) allows for the extraction of the pooled features with complementary information. The corresponding outputs (i.e., the channel weight from the two pooled feature sets) are then aggregated through a summation operation and subjected to a sigmoid activation layer, where the judgment regarding “attention” or “no attention” is made. After that, the output from the skip pathway of the densely connected module is multiplied channel-wise with the computed weights to give higher attention to more relevant feature channels. In the next step, the channel-weighted features are passed to the *spatial-attention* mechanism, where a max-pooling layer is used to identify the most informative spatial areas of the feature channels, and an average pooling layer is adopted to capture the overall information within that area. Finally, a 5×5 convolutional layer with a sigmoid activation function is utilized to generate the spatial attention weights for the input feature channels.

The entire process, performed as part of the proposed mixed attention mechanism, is formally expressed as follows:

$$\begin{gathered}M_c = \sigma_2\left(MLP\left(AvgPool\left(H\right)\right) + MLP\left(MaxPool\left(H\right)\right)\right) \\ = \sigma_2\left(\omega_1\left(\omega_0\left(H_{avg}^c\right)\right) + \omega_1\left(\omega_0\left(H_{\max}^c\right)\right)\right), \\ \omega_0 \in \mathbb{R}^{\frac{C}{r}\times C}, \omega_1 \in \mathbb{R}^{C\times\frac{C}{r}}\end{gathered} \tag{5}$$

where M_c are the channel attention weights for the input features, H denotes the feature from skip pathway of the densely connected extractor, $\sigma_2(\cdot)$ stands for the sigmoid activation function, $MLP(\cdot)$ represents the multi-layer perceptron with the $LeakReLU$ activation function, and r indicates a hyper-parameter corresponding to the dimensionality reduction rate. The computed weights are used in the channel-attention mechanism as follows:

$$H' = H \otimes M_c, \tag{6}$$

where $\otimes$ denotes the Hadamard product, and H' represents the channel-weighted image features.

The spatial-attention mechanism is described as follows:

$$\begin{gathered}M_s = \sigma_2\left(Conv^{5\times5}\left(\left[AvgPool\left(H'\right), MaxPool\left(H'\right)\right]\right)\right) \\ = \sigma_2\left(Conv^{5\times5}\left(\left(H'\right)_{avg}^s, \left(H'\right)_{\max}^s\right)\right),\end{gathered} \tag{7}$$

where M_s are the spatial attention weights for the (channel-weighted) input features, $\sigma_2(\cdot)$ represents the *sigmoid* activation function, and $Conv^{5\times5}(\cdot)$ denotes the convolution operation with the convolution kernel of size 5×5. The calculated spatial-attention weights are then applied as follows:

$$\hat{H} = H' \otimes M_s, \tag{8}$$

where (8) shows that the spatial attention weights are multiplied element-wise with channel-weighted input features to obtain the final mixed-attention feature tensor $\hat{H}$.

3.4. Emotion Classification

After the feature extraction stage, the spatial dimensions of the attention-weighted features $\hat{H}$ are transformed to a spatial resolution of 7×7, flattened, and passed into a fully-connected layer. This layer captures (long-range) dependencies between individual parts of the feature maps and facilitates the classification process, and can be expressed as:

$$\hat{y} = \sigma_3\left(\omega \cdot \hat{H} + b\right), \tag{9}$$

where $\hat{y}$ represents the probability distribution over the target classes, ω, and b denote the weight matrix and bias of the fully connected layer, respectively, and $\sigma_3(\cdot)$ stands for the $Softmax$ activation function.

In order to train and optimize the whole model, a standard cross-entropy loss is selected, i.e.:

$$Loss\left(y, \hat{y}\right) = -\frac{1}{N}\sum_{k\in N} y_k \log \hat{y}, \tag{10}$$

where N represents the number of samples, and y and $\hat{y}$ stand for the true probability distribution and predicted probability distribution of the k^{th} sample, respectively.

4. Experimental Evaluation

In this section, we report experimental results that show in comparative experiments with several state-of-the-art methods from the literature: (*i*) the performance of the proposed model in a compound setting with a mixed dataset composed of three popular micro-expression datasets, (*ii*) the performance of the model on each of the three datasets separately, (*iii*) comprehensive ablation experiments that demonstrate the impact of the proposed attention mechanism, (*iv*) the impact of the densely-connected feature extractor on the overall MER performance, and (*v*) the (time and space) complexity of the model.

4.1. Experimental Datasets

Two types of datasets are used in the experiments, i.e., macro-expression datasets for model pretraining, and micro-expression datasets for fine-tuning and testing.

Table 1: Characteristics of the selected datasets

Dataset	Type		Number			Resolution		Frame rate
	Dataset	Samples	Subjects	Samples	Emotions	Samples	Face	
CK+ [47]	In-lab	Image	123	593	7	-	-	-
RAF-DB [48]	In-the-wild	Image	-	29672	7	-	-	-
FER-2013 [49]	In-the-wild	Image	-	35886	7	48 × 48	-	-
CASME II [52]	In-lab	Video, image	26	255	7	640 × 480	280 × 340	200fps
SAMM [53]	In-lab	Video, image	26	159	7	2040 × 1080	400 × 400	200fps
SMIC-HS [30]	In-lab	Video, image	16	164	7	640 × 480	190 × 230	100fps

4.1.1. Selection of the Datasets

Three popular macro-expression datasets were selected to estimate the initial parameters for the feature-extraction model, i.e., the CK+ [47], the RAF-DB [48], and the FER-2013 dataset [49]. Details on these three datasets are shown in Table 1. As can be seen, two of the selected datasets were captured in uncontrolled (also called *in-the-wild*) settings and one was acquired in a laboratory environment. This composition of datasets was chosen to ensure the pretraining model performs well across a diverse set of acquisition conditions. For a similar reason, we also select three micro-expression datasets, i.e., CASME II [52], SAMM [53], and SMIC [30]. Detailed characteristics of these three datasets are also listed in Table 1.

Following prior work in this area, we address a three-class MER problem in this paper and partition the samples in the macro- and micro-expression datasets into three categories, namely, *negative*, *positive*, and *surprise*. Because of this categorization, some of the facial expressions present in the datasets are excluded from the experimentation. The final categorization (that also allows the mixing of the datasets during the experiments) is shown in Table 2.

Table 2: Organization of the datasets for the experiments

Dataset	Categorization			Excluded
	Negative	Positive	Surprise	
CK+ [47]	Anger, Contempt Disgus, Fear Sadness	Happiness	Surprise	-
RAF-DB [48]	Fear, Disgust Sadness Anger	Happiness	Surprise	Neutral
FER-2013 [49]	Anger, Fear Disgust Sadness	Happiness	Surprise	Neutral
CASME II [52]	Disgust, Repression	Happiness	Surprise	Others Fear Anger
SAMM [53]	Anger, Fear Disgust Contempt	Happiness	Surprise	Others
SMIC-HS [30]	Negative	Positive	Surprise	-

4.1.2. Processing of the Datasets

After the dataset partitioning, the micro-expression datasets are processed by the EVM algorithm. Through preliminary experiments, we found that the movement frequency of micro-expression changes very slightly, and its movement is roughly in the frequency range of $0.1 - 0.4$ Hz, so this range is targeted during the magnification. To achieve a good trade-off between noise amplification the micro-expression-motion amplification, we chose the amplification factor of 15 in the experiments. For each input video, the apex frame(s) are extracted and the face region is cropped and rescaled to the fixed size of 112×112 pixels by the functionality offered by OpenCV's face recognition algorithm [54].

4.1.3. Pretraining

In the pretraining process, the source domain (macro-expression) datasets are given as input to the model for training. The specific hyper-parameters used during this stage are shown in Table 3. During the

Table 3: Hyper-parameter configuration for pretraining

Hyper-parameter	Value
Batch size	16
# Epochs	200
Learning rate lr	0.01
lr reduction factor	0.1
Patience	20

pretraining process, 20% of the samples in the training set are used for validation. To ensure the generalization and effectiveness of the pretraining model, we maintain the balance of the sample size in the three categories (negative, positive, and surprise), and the weights corresponding to the highest accuracy on the validation set are considered the optimal weights computed during the pretraining. The accuracy on the validation set reached 87.15% during the best pretraining run.

4.2. Implementation Details

The model is implemented using the Keras deep learning framework. The pretraining and training of the feature extraction model are conducted on a 64-bit UBUNTU16.04 system with an E5-2598V4 CPU, and four NVIDIA TeslaV100 GPUs. Table 4 lists the detailed configuration for training (fine tuning) in the target (micro-expression) domain. The *Adam* optimizer is used in the training process. To prevent overfitting, an ’early stopping’ criterion is also implemented and training is interrupted if the performance on the validation set fails to improve for 20 consecutive epochs. To ensure good model convergence, the learning rate is adaptively reduced as the training progresses.

Table 4: Hyper-parameters configuration for the final training

Hyper-parameter	Value
Batch size	32
Epochs	200
Learning rate lr	0.01
lr reduction factor	0.1
Patience	20

4.3. Performance Measures

In accordance with the standard evaluation methodology [55], we use accuracy, unweighted $F1_{\text{score}}$ ($UF1$), and unweighted average recall rate (UAR) as the performance indicators in our experiments. Accuracy (Acc) is defined as the fraction of correctly classified samples with respect to the total number of samples. $UF1$ represents the average value of the $F1_{\text{score}}$ of each class, that is:

$$UF1 = \frac{\sum_{k=1}^{m} F_k}{m}, \tag{11}$$

where m denotes the number of classes, and F_k is the $F1_{\text{score}}$ of the k^{th} class. F_k is given by the harmonic mean between precision (P_k) and recall (R_k) of the k^{th} class, that is,

$$F1_{\text{score}} = 2 \times \frac{P_k \times R_k}{(P_k + R_k)}, \tag{12}$$

$$P_k = \frac{TP_k}{(TP_k + FP_k)}, R_k = \frac{TP_k}{(TP_k + FN_k)}, \tag{13}$$

where TP_k, FP_k, and FN_k denote the number of true positives, false positives, and false negatives for the k^{th} class, respectively.

Finally, UAR is given by:

$$UAR = \frac{\sum_{k=1}^{m} TP_k}{m}, \tag{14}$$

where TP_k and m denote the number of true positives for the k^{th} class and the number of classes, respectively.

4.4. Experimental Setting

We define two benchmarks to evaluate the proposed method, i.e., the composite-dataset evaluation (CDE) benchmark and the single-dataset evaluation (SDE) benchmark. In the CDE benchmark, we use samples from *all experimental datasets* to fine-tune the proposed model for micro-expression recognition. The model, therefore, sees a diverse set of image characteristics during training and is expected to generalize better to a wide variety of input samples. Different test datasets are then utilized with the CDE benchmark. The SDE benchmark, on the other hand, corresponds to the traditional experimental setup used when evaluating MER models. With this benchmark, the evaluated MER model is fine-tuned and tested on a single dataset at the time. The reported performance, therefore, reflects the performance of the evaluated model in a specific scenario/setting and points to its generalization capabilities across different data characteristics.

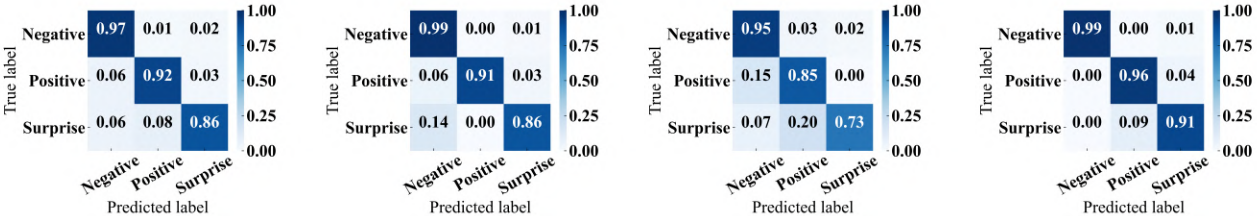
We select the leave-one-subject-out (LOSO) evaluation protocol for all experiments regardless of the benchmarking methodology used (CDE or SDE). The LOSO protocol represents the standard and most widely used protocol in MER research, where each subject is excluded from the training procedure once, and performance is then reported through aggregate statistics over all (unseen/excluded) subjects.

4.5. Comparative Results

In this section, we report comparative results with state-of-the-art methods from the literature to demonstrate the superiority of the proposed approach over its competitors. Additionally, we also analyze the main causes of misclassification.

4.5.1. Experiments Under the CDE Benchmark

Under the CDE benchmark, we compare our approach against the following state-of-the-art competitors: LBP-TOP [8], Bi-WOOF [11], CapsuleNet [41], OFF-ApexNet [42], Dual-Inception [56], STSTNet [16], EMR [27], RCN [17], ICE-GAN [57], and MERASTC [58]. Experimental results are reported for the three micro-expression datasets separately (i.e., CASME II, SAMM, and SMIC-HS), but also for a combined dataset that includes samples from all three test datasets. We refer to this combined dataset as *Mixed* hereafter. All tested models are fine-tuned using data from the three experimental micro-expression datasets and tested with the LOSO protocol.



(a) Mix dataset (b) CASME II dataset (c) SAMM dataset (d) SMIC-HS dataset

Figure 5: Confusion matrices generated by the proposed approach for the Mixed, CASME II, SAMM, and SMIC-HS datasets under the CDE benchmark. Micro-expression recognition (MER) is conducted on a three-class problem, where samples are classified into the positive, negative or surprise class. The figure is best viewed in color.

Experimental results on the Mixed dataset. We first discuss and analyze the experimental results obtained on the mixed dataset. As shown by the confusion matrix in Figure. 5 (a), the classification of *negative* expressions is the most accurate, followed by that of *positive* expressions, whereas the classification accuracy of *surprise* is the lowest. The main reason for this result is the uneven distribution of the number of samples in the datasets. There are 252 samples in the *negative* category, 109 samples in the *positive* category, and only 86 samples in the *surprise* category. This data distribution is also reflected in the recognition results and introduces a preference in the model toward the more represented expression categories.

When comparing the proposed approach to the competing handcrafted feature-extraction methods and the more recent deep-learning techniques in Table 5, we see a clear advantage for our approach. In addition to the convincing classification accuracy of 93.74%, the other two performance indices $UF1$ and UAR also clearly point to the superiority of the proposed approach, i.e., $UF1 = 0.9213$ and $UAR = 0.9167$.

Compared with the competing deep-learning methods, our approach performs better in solving the MER classification problem. The results for OFF-ApexNet [42] (which introduced the concept of apex frames and greatly simplified the MER recognition process) show that just using apex frames is insufficient for MER, because compared to macro-expression, the expression intensity in the apex frames is still too weak for micro-expressions. Therefore, the proposed method improves by 0.2017 and 0.2071 with respect to the $UF1$ and UAR scores, respectively, over OFF-ApexNet. Furthermore, OFF-ApexNet, CapsuleNet [41], Dual-Inception [56], STSTNet [16], and RCN [17] all adopt shallow networks to overcome the problem of small numbers of samples in the micro-expression datasets. However, shallow models are only able to extract features with limited expressive power, leading to suboptimal performance. Conversely, if the model is too deep and the number of samples in the micro-expression datasets is too small, the model easily overfits and results in poor generalization ability. While the above-mentioned models suffer from such shortcomings, the proposed approach overcomes these issues and shows clear improvements over the competitors.

Table 5: Performance comparison with the competing techniques on the Mixed dataset under the CDE benchmark.

Handcrafted methods	UAR	$UF1$	*Acc* (%)
Bi-WOOF [11]	0.6227	0.6296	-
LBP-TOP [8]	0.5785	0.5882	-
Deep learning methods	UAR	$UF1$	*Acc* (%)
CapsuleNet [41]	0.6506	0.6520	-
OFF-ApexNet [42]	0.7096	0.7196	-
Dual-Inception [56]	0.7278	0.7322	-
STSTNet [16]	0.7605	0.7353	-
EMR [27]	0.7824	0.7885	-
SA-AT [28]	0.5958	0.5936	-
RCN [17]	0.7165	0.7052	-
ICE-GAN [57]	0.8410	0.8450	-
MERASTC [58]	0.9160	0.9200	-
Ours	**0.9167**	**0.9213**	**93.74**

When looking at the results for the EMR [27] method that also uses EVM to increase the intensity of facial expressions, but relies on ResNet features for data representation, we again see that our solution has a clear edge. This edge is ensured by the densely connected module and our mixed attention mechanism, which leads to highly discriminative features for MER and makes better use of the complete information contained in the input data. As a result of the proposed design, our method improves by 0.1328 in terms of $UF1$ and by 0.1343 in terms of the UAR score on EMR.

In the SA-AT [28] approach, a ResNet backbone, and an attention mechanism are combined, and a transfer learning method is adopted to fine-tune the model for MER. However, the performance of this method is not ideal. Compared to our method, the correspondence between samples in the source and target domain is insufficient, which leads to poor results during transfer learning. Another weak point is that the feature extraction model of SA-AT, ResNet, introduces more parameters than the densely connected module designed in this paper, thus resulting in limited generalization capabilities across image characteristics. Additionally, the standard sequential feed-forward model topology leads to a loss of potentially important image information. Thus, in terms of experimental results, our method yields 0.3209 and 0.3277 higher performance scores than SA-AT. ICE-GAN [57] uses a GAN network to generate new data to enrich the datasets, which is shown to be beneficial for performance, but the newly generated data needs an appropriate feature extraction model to match it. MERASTC [58] extracts features around key facial points, which greatly improves the generalization of the model. However, the traditional convolutional layer stack model is chosen as the feature extraction model, which adversely affects its performance. As a result of these issues, our method also outperforms these two approaches.

The proposed method also compares favorably when compared with the traditional handcrafted feature-extraction techniques [8, 11]. The reason for this outcome is that handcrafted features are not specialized enough for the extraction of representative facial features for MER. These features are too sensitive with respect to environmental factors and are, hence, not suitable for images captured in complex environments. In contrast, the features extracted by our method are more targeted and have better robustness to environmental factors. Therefore, the two performance indicators, $UF1$ and UAR, for our method are 0.2917 and 0.2940 higher than for the traditional method with the best performance, respectively.

Experimental results on the CASME II, SAMM, and SMIC-HS datasets. To better analyze the performance of our model, experiments are also carried out on the CASME II, SAMM, and SMIC-HS datasets separately. Figures. 5 (b), (c), and (d) show the confusion matrices obtained from the experiments on these datasets. It can be seen that the proposed method has good classification performance on the CASME II and SMIC-HS datasets, but due to the unbalanced distribution of data samples across the classes in these datasets, the classification accuracy for *surprise* is not as good as for *negative* and *positive* expressions. The classification accuracy on the CASME II and SMIC datasets is higher than the classification accuracy on the SAMM dataset. Considering the confusion matrix on this latter dataset, it appears that most misclassifications occur between *positive* expressions and *surprise*. There are two reasons for this result. The first one is that in the SAMM dataset, the number of surprise samples is only 30, which is far less than the number of *positive* and *negative* samples. The second reason is that all the samples in the SAMM dataset are black-and-white images, while all the samples in the CASME II and SMIC-HS datasets are color images. The mixed attention module in our feature extraction model is more sensitive to the channel features and spatial features of color images. Therefore, the final classification performance on the CASME II and SMIC-HS datasets is significantly better than that observed for the SAMM dataset.

Next, we compare our method with the baselines that were also considered in the previous section. The detailed results of the comparison are shown in Table 6. For the *CASME II dataset*, both the handcrafted feature methods and deep learning methods have relatively good performance. The reason is that the frame rate of the CASME II dataset is higher, and the resolution of each frame image is also higher than with the other two datasets. Therefore, the recognition performance on this dataset is relatively high for all tested techniques. Nonetheless, our method improves the performance of the competing methods to a certain degree and results in the final accuracy of 94.67%. On the *SAMM dataset*, compared to the traditional methods exploiting handcrafted features, our method is 0.3285 and 0.3274 higher in terms of the considered performance indices. We ascribe this result to the fact that traditional handcrafted features do not have strong adaptability to complex environmental changes. With our method, on the other hand, different scenes and different types of images are used during pretraining, which enables the model to have strong adaptability to the micro-expression features in this dataset. Our method is also superior to most previous deep learning methods in terms of performance indicators on this dataset. However, two methods (ICE-GAN and MERASTC) perform slightly better than our method, because the images extracted in the SAMM dataset are black and white, which makes the channel features of the images lack significance. For the

Table 6: Performance comparison with competing methods on the CASME II, SAMM, SMIC-HS datasets under the CDE benchmark.

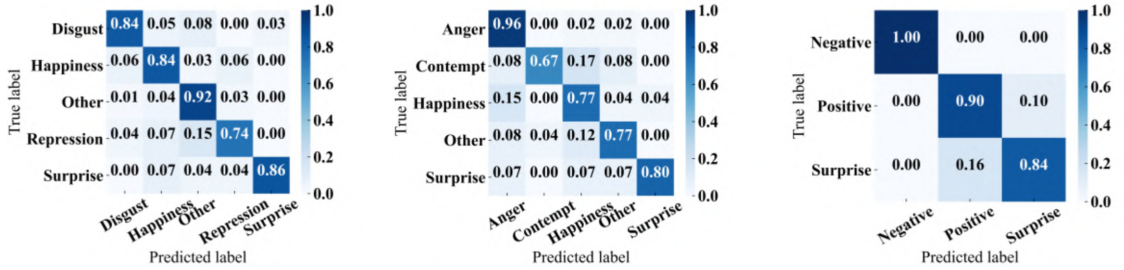
Handcrafted Method	CASME II			SAMM			SMIC-HS		
	UAR	$UF1$	Acc (%)	UAR	$UF1$	Acc (%)	UAR	$UF1$	Acc (%)
Bi-WOOF [11]	0.8026	0.7805	-	0.5139	0.5211	-	0.5829	0.5727	-
LBP-TOP [8]	0.7429	0.7026	-	0.4102	0.3954	-	0.5280	0.2000	-
Deep learning method	UAR	$UF1$	Acc (%)	UAR	$UF1$	Acc (%)	UAR	$UF1$	Acc (%)
CapsuleNet [41]	0.7018	0.7068	-	0.5989	0.6209	-	0.5877	0.5820	-
OFF-ApexNet [42]	0.8681	0.8764	-	0.5392	0.5409	-	0.6695	0.6817	-
Dual-Inception [56]	0.8560	0.8621	-	0.5663	0.5868	-	0.6726	0.6645	-
STSTNet [16]	0.8686	0.8382	-	0.6810	0.6588	-	0.7013	0.6801	-
EMR [27]	0.8209	0.8293	-	0.7152	0.7754	-	0.7530	0.7461	-
SA-AT [28]	0.7552	0.7607	-	0.4868	0.4476	-	0.5463	0.5512	-
RCN [17]	0.8563	0.8087	-	0.6976	0.6771	-	0.5991	0.5981	-
ICE-GAN [57]	0.8680	0.8760	-	0.8230	0.8550	-	0.7910	0.7900	-
MERASTC [58]	0.9500	0.9330	-	0.8460	0.8300	-	0.8620	0.7900	-
Ours	**0.9174**	**0.9340**	**94.67**	**0.8415**	**0.8485**	**90.15**	**0.9512**	**0.9510**	**95.76**

SMIC-HS dataset, our proposed method significantly outperforms all previously proposed methods. On this dataset, the performance of the competing methods is always the worst across all three datasets. However, the proposed feature extraction model can extract highly informative features for MER and results in the overall accuracy of 95.76% on the SMIC-HS, improving significantly on the runner-up, the MERASTC approach. Overall, the presented results clearly show that our model clearly shows a better generalization ability and stronger robustness to the environmental factors than the competing approaches.

The Student's t-test results under the CDE benchmark. Given that statistical characteristics reflect the robustness of the model, we have substantiated the robustness of the proposed model using the Student's t-test under the CDE benchmark. The p-values were instrumental in gauging the robustness disparity between our method and existing methods. In terms of the UAR metric, the p-values of our method when juxtaposed with those of Bi-WOOF [11], LBP-TOP [8], CapsuleNet [41], OFF-ApexNet [42], Dual-Inception [56], STSTNet [16], EMR [27], SA-AT [28], RCN [17], ICE-GAN [57], and MERASTC [58] are 0.0079, 0.0052, 0.0017, 0.017, 0.0145, 0.0166, 0.0037, 0.0047, 0.0269, 0.0439, and 0.3255, respectively. For the $UF1$ metric, the p-values corresponding to our method and the above methods are 0.0047, 0.0147, 0.0019, 0.0160, 0.0123, 0.0070, 0.0098, 0.0045, 0.0108, 0.0638, and 0.1627, respectively. Notably, aside from MERASTC for UAR and $UF1$ and ICE-GAN for $UF1$, all other methods show p-values less than 0.05 when compared to our proposed method, indicating a stronger robustness of our approach. Although MERASTC exhibits the closest robustness to our method, our approach is significantly more straightforward to implement.

4.5.2. Experiments under the SDE benchmark

Under the SDE benchmark, we compare our method against the following competing techniques: LBP-TOP + AdaBoost [59], STCLQP [9], FDM [60], LBP-TOP [61], AlexNet [62], ELRCN [63], SSSN [64], DSSN [64], and TSCNN [45]. We analyze the experimental results separately for each of the three datasets (CASEM II, SAMM, and SMIC-HS datasets). For each experiment, a single dataset is used for fine-tuning and testing using the LOSO protocol.



(a) CASME II dataset (b) SAMM dataset (c) SMIC-HS dataset

Figure 6: Confusion matrices generated by the proposed approach for the CASME II, SAMM, and SMIC-HS datasets under the SDE benchmark. Micro-expression recognition (MER) on the CASME II and SAMM datasets is for the five-class problem, whereas MER on the SMIC-HS dataset is for the three-class problem. The figure is best viewed in color.

Experimental results on the CASME II dataset. We consider a 5-class recognition problem on the CASME II dataset to facilitate comparison with prior work, where the target emotional categories are *disgust*, *happiness*, *repression*, *surprise*, and *other*. As shown in Figure 6 (a), the *other* category has the highest recognition accuracy, with an accuracy of 92%. *Repression*, on the other hand, exhibits the lowest performance with a recognition accuracy of only 74%. There are two main reasons for this situation, one being that the *other* class has the largest number of samples (99), whereas *repression* has the smallest number of samples (27). Secondly, it can be observed that 15% of the samples in the *repression* category are misclassified as *other*, indicating that the samples in the *repression* category have similar emotional expressions as some of the samples labeled as *other*.

Table 7 illustrates the performance of the proposed method compared to the selected state-of-the-art MER methods. As can be seen from the results, the performance of the proposed method is significantly stronger than that of the competing methods. The best-performing MER method based on handcrafted features is STCLQP, with $UF1$ and Acc reaching values of 0.5836 and 58.39%, respectively. Among the deep learning methods (AlexNet, ELRCN, SSSN, DSSN, TSCNN), TSCNN yields the best performance. The method is a three-stream MER approach that considers the global apex frame image, the optical flow map, and the local apex frame image for the recognition process. Because the model uses information-rich representations of the facial emotions it has excellent recognition performance. Compared to the best-performing handcrafted method STCLQP, TSCNN achieves a 0.2634 higher $UF1$ score and a 27.96% higher Acc score. Our method uses a transfer-learning strategy to overcome the problem of limited sample size, while

Table 7: Performance comparison with the existing work on CASME II dataset under SDE benchmark.

Handcrafted methods	$UF1$	Acc (%)
LBP-TOP + AdaBoost [59]	0.3337	43.78
STCLQP [9]	0.5836	58.39
FDM [60]	0.4700	41.96
LBP-TOP [61]	0.4700	51.00
Deep learning methods	$UF1$	Acc (%)
AlexNet [62]	0.6675	62.96
ELRCN [63]	0.5000	52.44
SSSN [64]	0.7151	71.19
DSSN [64]	0.7297	70.78
TSCNN [45]	0.8070	80.97
Ours	**0.8408**	**86.35**

the recognition performance is also improved with the help of macro-expressions pretraining. Furthermore, the densely connected structure and the mixed attention mechanism lead to highly discriminative features, which in turn enable the proposed method to clearly outperform the state-of-the-art in terms of recognition performance, with 86.35% on *Acc* and 0.8408 on $UF1$.

Experimental results on the SAMM dataset. Similar to the experiments on the CASME II dataset, a five-class recognition task is considered on the SAMM dataset to enable comparison with existing methods. The emotions included in the SAMM dataset are *anger*, *contempt*, *happiness*, *surprise*, and *other*. From the confusion matrix presented in Figure 6 (b), it can be found that the lowest recognition accuracy of 67% is achieved for the samples of the *contempt* class. It is interesting to note that 17% of the *contempt* samples were misclassified as belonging to *happiness*. Since both *contempt* and *happiness* are accompanied by a rise in the corners of the mouth, the two emotions are very easily confused in MER. Moreover, happiness contains approximately twice as many samples as contempt, which is another reason for the misclassification.

Unlike the CASME II dataset, the SAMM dataset contains samples of grey-scale images, and the reduction of channel information in the samples increases the difficulty of MER. Therefore, both the handcrafted MER methods and the deep learning MER methods have decreased performance on the SAMM dataset. From Table 8, the worst performance of these state-of-the-art methods is seen with LBP-TOP, while the best performance is observed with TSCNN, where TSCNN outperforms LBP-TOP by 49.88% and 0.5286 in terms of *Acc* and $UF1$ scores, respectively. Although the recognition performance of our method on the SAMM dataset is degraded compared to the experimental results on the CASME II dataset, our approach still outperforms all of the tested state-of-the-art MER methods by a considerable margin.

Experimental results on the SMIC-HS dataset. Different from the CASME II and SAMM datasets, the experiments on the SMIC-HS dataset are conducted for a three-class task, where emotions include *negative*, *positive*, and *surprise*. Figure 6 (c) illustrates the confusion matrix corresponding to the experimental

Table 8: Performance comparison with the existing work on the SAMM dataset under the SDE benchmark.

Handcrafted methods	$UF1$	Acc (%)
LBP-TOP [8]	0.2892	34.56
LBP-SIP [31]	0.3133	36.03
HOG-TOP [32]	0.3403	36.06
HIGO-TOP [32]	0.3920	41.18
Deep learning methods	$UF1$	Acc (%)
AlexNet [62]	0.4260	52.94
SSSN [64]	0.4513	56.62
DSSN [64]	0.4644	57.35
TSCNN [45]	0.6942	71.76
Ours	**0.8178**	**84.44**

results of the proposed method on the SMIC-HS dataset. The samples in the *negative* category are all correctly classified, but there are still a small number of misclassifications that occur between *positive* and *surprise*. This is because there is a certain similarity in the way these two emotions are expressed.

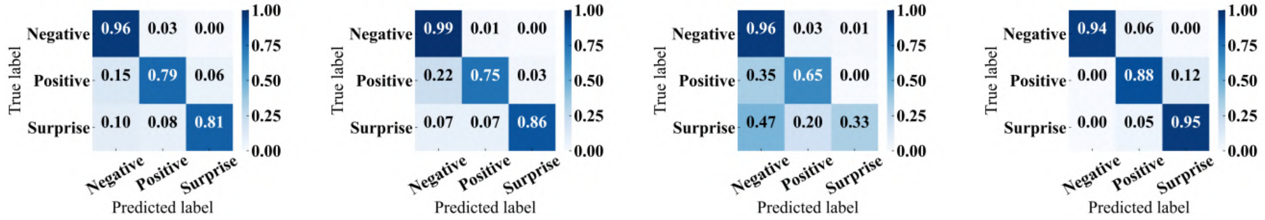
Table 9: Performance comparison with the existing work on the SMIC-HS dataset under the SDE benchmark.

Handcrafted methods	$UF1$	Acc (%)
LBP-TOP + AdaBoost [59]	0.4731	44.34
STCLQP [9]	0.6381	64.02
FDM [60]	0.5380	54.88
Bi-WOOF + Phase [65]	0.6730	68.29
Deep learning methods	$UF1$	Acc (%)
AlexNet [62]	0.6013	59.76
GoogleNet [66]	0.5511	51.23
OFF-ApexNet [42]	0.6709	67.68
SSSN [64]	0.6329	63.41
DSSN [64]	0.6462	63.41
TSCNN [45]	0.7236	72.74
Ours	**0.9139**	**92.73**

As shown in Table 9, Bi-WOOF+Phase resulted in $UF1$ and Acc scores of 0.6730 and 68.29%, respectively, which is the highest among the handcrafted MER methods. Furthermore, Bi-WOOF+Phase also performed better than some of the deep learning methods (e.g., AlexNet, GoogleNet, OFF-ApexNet, SSSN, and DSSN). Among all of the state-of-the-art methods considered, TSCNN showed the best performance. However, the method proposed in this paper significantly outperforms TSCNN both in terms of $UF1$ as well as Acc scores. Specifically, our approach yields an $UF1$ score that is 0.1903 higher than that of TSCNN and an Acc score

Table 10: Results of the ablation experiment under the CDE benchmark.

Method	Mixed			CASME II			SAMM			SMIC-HS		
	UAR	*UF*1	*Acc* (%)	*UAR*	*UF*1	*Acc* (%)	*UAR*	*UF*1	*Acc* (%)	*UAR*	*UF*1	*Acc* (%)
Ours	0.9167	0.9213	93.74	0.9174	0.9340	94.67	0.8415	0.8485	90.15	0.9512	0.9510	95.76
w/o Attention	0.8557	0.8691	89.26	0.8653	0.8887	91.33	0.6477	0.6890	82.58	0.9265	0.9215	92.73



(a) Mixed dataset (b) CASME II dataset (c) SAMM dataset (d) SMIC-HS dataset

Figure 7: Confusion matrices of the ablation experiment on different datasets under the CDE benchmark. The matrices show the performance of the proposed approach implemented without the mixed attention module. The reference results with the attention mechanism are shown in Figure. 5. The figure is best viewed in color.

that is 19.99% higher, which clearly speaks of the capabilities of our approach on the SMIC-HS dataset.

4.6. Ablation Study

One of the key components of the model proposed in this paper is the mixed attention mechanism that combines spatial attention and channel attention to facilitate the extraction of descriptive image features. To demonstrate the impact of the proposed mixed attention mechanism, we design a comprehensive two-stage ablation study. In the first stage, we remove the entire attention mechanism from our model and observe results, whereas in the second stage, we ablate each of the steps of the proposed mechanism and analyze its impact on the overall MER performance. Additionally, we also report qualitative/visual results. All experiments of the ablation study are conducted in accordance with the CDE benchmark and LOSO experimental protocol.

Mixed Attention Ablation. The results of the first-stage ablation experiments are presented in Table 10 and Figure. 7. Note that the confusion matrices in Figure. 7 only report results for the proposed model without the attention mechanism, while the results for the complete model are given in Figure. 5 As can be seen from the presented results, in case the mixed attention module is absent, the accuracy on the mixed dataset clearly decreases. The performance of the model with the mixed attention module is 4.48%, 0.522, and 0.61 higher, with respect to the *Acc*, $UF1$, and the UAR score on this dataset, respectively. This shows that under the compound conditions, the attention mechanism makes the feature extraction model more targeted towards discriminative image features and results in significant performance improvements compared to the setting where no attention is used.

It can also be observed from Figures. 7 (b), (c), and (d) that misclassified samples stem mainly from the

SAMM and the SMIC-HS datasets. On the SAMM dataset, most *surprise* expressions are misclassified as *negative*, indicating that the mixed attention module does indirectly enhance the characteristics of key areas and improves the accuracy of MER on the SAMM dataset. On the SMIC-HS dataset, misclassifications mainly occur between *surprise* and *positive*. Some samples whose emotional polarity is *positive* are misclassified as *surprise*. This is because the two emotional polarities have certain similarities in expression state. On the CASME II dataset, the attention also results in significant performance improvements as on the other two datasets. Overall, the reported results suggest that the proposed mixed attention module is indispensable for efficient feature extraction with the proposed model.

Fine–Grained Ablations. To further explore the effectiveness of the mixed attention mechanism, we perform a fine-grained ablation study and explore the impact of the individual components of the proposed mechanism on MER performance. Specifically, we consider four settings, i.e., no attention, only channel attention, only spatial attention, and mixed attention. The results of this experiment are presented in Table 11 and in the form of Grad-CAM [67] visualizations in Figure. 8.

As can be seen, the feature extraction model without the attention mechanism focuses on a broad spatial area within the facial images. Due to the insufficient sensitivity to the local information of the face, the performance of the model is not ideal.

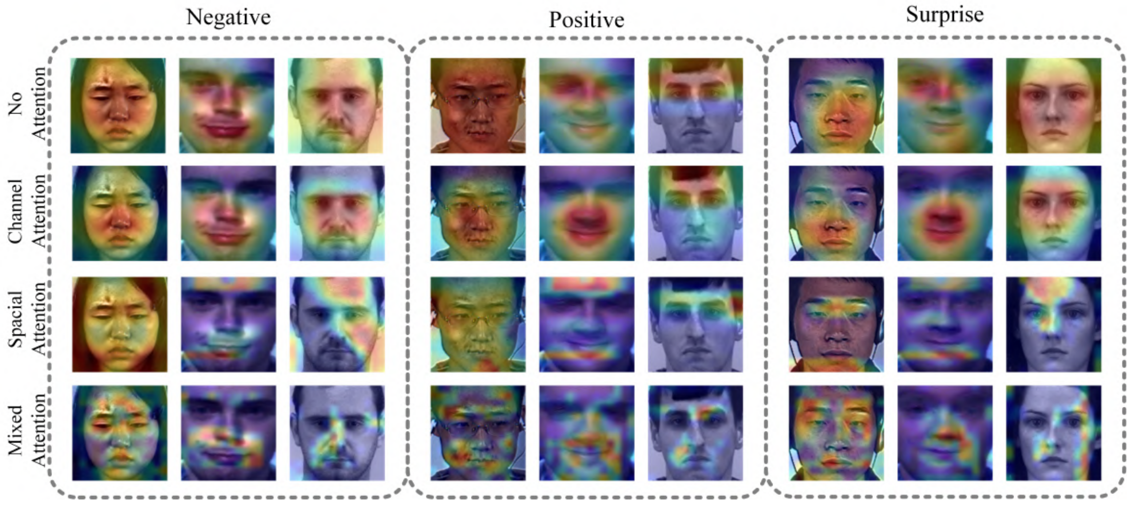


Figure 8: Illustration of the impact of the attention mechanism in terms of Grad-CAM visualizations. The heat maps are presented for four different model configurations, i.e., without the attention mechanism, using only the channel part of the attention mechanism, using only the spatial part of the attention mechanism, and using the complete mixed attention mechanism. The examples show that the mixed attention mechanism yields superior performance.

Moreover, when only the spatial attention mechanism or the channel attention mechanism is used, the ability of the model to focus on local image information is significantly improved. Since the channel attention mechanism and the spatial attention mechanism capture the local information from two different dimensions of the feature image, there are differences in the selection of local regions. The mixed attention mechanism combines the characteristics of the two individual attention mechanisms.

Table 11: Comparison of attention mechanisms under the CDE benchmark.

Method	Mixed			CASME II			SAMM			SMIC-HS		
	UAR	$UF1$	*Acc* (%)	UAR	$UF1$	*Acc* (%)	UAR	$UF1$	*Acc* (%)	UAR	$UF1$	*Acc* (%)
No attention	0.8557	0.8691	89.26	0.8653	0.8887	91.33	0.6477	0.6890	82.58	0.9265	0.9215	92.73
Channel attention only	0.8892	0.8928	91.28	0.8952	0.8957	91.33	0.7874	0.7931	88.64	0.9245	0.9256	93.33
Spatial attention only	0.8766	0.8942	91.05	0.9108	0.9121	92.67	0.6644	0.7164	84.09	0.9428	0.9485	95.15
Mixed attention (proposed)	**0.9167**	**0.9213**	**93.74**	**0.9174**	**0.9340**	**94.67**	**0.8415**	**0.8485**	**90.15**	**0.9512**	**0.9510**	**95.76**

Table 12: Comparison of the densely-connected and residual feature extractors under the CDE benchmark.

Method	Mixed			CASME II			SAMM			SMIC-HS		
	UAR	$UF1$	*Acc* (%)	UAR	$UF1$	*Acc* (%)	UAR	$UF1$	*Acc* (%)	UAR	$UF1$	*Acc* (%)
Residual	0.7846	0.7954	84.56	0.8163	0.8317	86.67	0.6507	0.6426	83.33	0.8102	0.8113	83.64
Densely connected (proposed)	**0.9167**	**0.9213**	**93.74**	**0.9174**	**0.9340**	**94.67**	**0.8415**	**0.8485**	**90.15**	**0.9512**	**0.9510**	**95.76**

The attention to the effective area is not only further improved, but can also help to focus the feature extraction process on the most important sources of information. This process is illustrated with the heat maps in Figure. 8. Meanwhile, the distribution of the attention heatmap can also well demonstrate the sparse emotional distribution of micro-expressions.

4.7. Impact of Feature Extractor Design

The feature extraction stage of the proposed MER approach is based on a densely connected module. With the help of the densely connected structure, the feature loss is effectively controlled, and the integrity of the output feature is guaranteed. To demonstrate the impact of the densely connected structure on MER performance, we replace the densely-connected blocks with residual blocks and then conduct comparative experiments under the CDE benchmark on the mixed dataset and with the proposed attention mechanism.

As shown in Table 12, the densely-connected structure performs better than the residual structure. The reason for this result can be ascribed to the fact that the output features of each layer of the densely connected blocks are used as the input feature of the subsequent feature extraction layers during the feature extraction process. This structural characteristic significantly reduces the loss of discriminative information along the network. The extracted features, thus, represent a powerful representation of the MER task. Moreover, the feature extraction model designed in this paper is observed to guarantee the integrity of the features and ensures that locally relevant information is used by combining the densely connected structure and the mixed attention mechanism.

4.8. Computational Cost

The time complexity of the model and the number of parameters are listed in Table 13. As can be seen, the model has a relatively low number of parameters compared to the standard deep learning model used in the (visual) recognition literature.

Table 13: FLOPs and number of parameters of our method

Method	Input size	FLOPs ($\times 10^6$)	Param (M)
Ours	112×112	6.26	3.115

5. Conclusion

At present, the existing methods for micro-expression recognition using video apex frames have no obvious pertinence to the image. However, in the method proposed in this paper, the preprocessing part converts the micro-expression features into an image form. Furthermore, a composite feature extraction model combining a densely-connected feature-extraction module and a mixed attention module is built according to the characteristics of the image. In the feature extraction process, the densely-connected feature-extraction module is used to reduce the loss of image features during processing. At the same time, the mixed attention module is used to process the channel characteristics and spatial characteristics of the image, which greatly improves the performance of single-frame micro-expression recognition. In addition, in the experimental part, the effectiveness of the model is demonstrated through experiments on multiple datasets, and the results show that the model has strong robustness in compound scenarios.

Future endeavors in micro-expression recognition must continue to address the challenges posed by the subtle and short duration of micro-expressions within video or image samples. Insights gleaned from transfer learning suggest that amplifying the motion intensity of micro-expressions themselves is not the only possible approach. Leveraging the prominence of macro-expressions to guide the extraction of micro-expression features can also enhance the precision of micro-expression recognition models. Moreover, given the restricted size of micro-expression datasets, there is a pressing need for models that can extract more comprehensive semantic cues from a limited amount of data and have a superior perception of facial emotion. Consequently, models that can allocate computational resources with greater accuracy are poised to demonstrate better performance in micro-expression recognition.

Acknowledgements

This work was supported in parts by the Chongqing Research Program of Basic Research and Frontier Technology (Grant No. cstc2021jcyj-msxmX0761), and the ARRS Research Programme P2-0250 (B) Metrology and Biometric Systems.

References

[1] M. Pantic, L. J. M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, IEEE Transactions on pattern analysis and machine intelligence 22 (12) (2000) 1424–1445.

[2] J. Xiao, C. Gan, Q. Zhu, Y. Zhu, G. Liu, Cfnet: Facial expression recognition via constraint fusion under multi-task joint learning network, Applied Soft Computing 141 (2023) 110312.

[3] C. Gan, L. Wang, Z. Zhang, Z. Wang, Sparse attention based separable dilated convolutional neural network for targeted sentiment analysis, Knowledge-Based Systems 188 (2020) 104827.

[4] P. Ekman, W. V. Friesen, Constants across cultures in the face and emotion., Journal of personality and social psychology 17 (2) (1971) 124.

[5] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, Y.-J. Liu, Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (9) (2021) 5826–5846.

[6] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (3) (2023) 2782–2800.

[7] P. Ekman, Lie catching and microexpressions, The philosophy of deception 1 (2) (2009) 5.

[8] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE transactions on pattern analysis and machine intelligence 29 (6) (2007) 915–928.

[9] X. Huang, G. Zhao, X. Hong, W. Zheng, M. Pietikäinen, Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns, Neurocomputing 175 (2016) 564–578.

[10] P. Ekman, Facial expression and emotion., American psychologist 48 (4) (1993) 384.

[11] S.-T. Liong, J. See, K. Wong, R. C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Processing: Image Communication 62 (2018) 82–92.

[12] K. Grm, W. J. Scheirer, V. Štruc, Face hallucination using cascaded super-resolution and identity priors, IEEE Transactions on Image Processing 29 (2020) 2150–2165.

[13] J. Deng, J. Guo, T. Liu, M. Gong, S. Zafeiriou, Sub-center arcface: Boosting face recognition by large-scale noisy web faces, in: European Conference on Computer Vision, Springer, 2020, pp. 741–757.

[14] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, V. Štruc, Privacy–enhancing face biometrics: A comprehensive survey, IEEE Transactions on Information Forensics and Security 16 (2021) 4147–4183.

[15] D. Patel, X. Hong, G. Zhao, Selective deep features for micro-expression recognition, in: 2016 23rd international conference on pattern recognition (ICPR), IEEE, 2016, pp. 2258–2263.

[16] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, Y.-C. Huang, Shallow triple stream three-dimensional cnn (STSTNet) for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[17] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, G. Zhao, Revealing the invisible with model and data shrinking for composite-database micro-expression recognition, IEEE Transactions on Image Processing 29 (2020) 8590–8605.

[18] Z. Zhai, J. Zhao, C. Long, W. Xu, S. He, H. Zhao, Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22086–22095.

[19] M. A. Takalkar, S. Thuseethan, S. Rajasegarar, Z. Chaczko, M. Xu, J. Yearwood, Lgattnet: Automatic micro-expression detection using dual-stream local and global attentions, Knowledge-Based Systems 212 (2021) 106566.

[20] R. Ni, B. Yang, X. Zhou, S. Song, X. Liu, Diverse local facial behaviors learning from enhanced expression flow for microexpression recognition, Knowledge-Based Systems 275 (2023) 110729.

[21] M. Peng, Z. Wu, Z. Zhang, T. Chen, From macro to micro expression recognition: Deep learning on small datasets using transfer learning, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 657–661.

[22] X. Jia, X. Ben, H. Yuan, K. Kpalma, W. Meng, Macro-to-micro transformation model for micro-expression recognition, Journal of Computational Science 25 (2018) 289–297.

[23] X. Ben, X. Jia, R. Yan, X. Zhang, W. Meng, Learning effective binary descriptors for micro-expression recognition transferred by macro-information, Pattern Recognition Letters 107 (2018) 50–58.

[24] Y. Zhang, Y. Jiang, J. Alireza, Mutual supervised fusion & transfer learning with interpretable linguistic meaning for social data analytics, ACM Transactions on Asian and Low-Resource Language Information Processing 22 (5) (2023) 1–20.

[25] Y. Zhang, Z. Zhou, W. Pan, H. Bai, W. Liu, L. Wang, C. Lin, Epilepsy signal recognition using online transfer tsk fuzzy classifier underlying classification error and joint distribution consensus regularization, IEEE/ACM transactions on computational biology and bioinformatics 18 (5) (2020) 1667–1678.

[26] Y. Jiang, Y. Zhang, C. Lin, D. Wu, C.-T. Lin, Eeg-based driver drowsiness estimation using an online multi-view and transfer tsk fuzzy system, IEEE Transactions on Intelligent Transportation Systems 22 (3) (2020) 1752–1764.

[27] Y. Liu, H. Du, L. Zheng, T. Gedeon, A neural micro-expression recognizer, in: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), IEEE, 2019, pp. 1–4.

[28] L. Zhou, Q. Mao, L. Xue, Cross-database micro-expression recognition: a style aggregated and attention transfer approach, in: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2019, pp. 102–107.

[29] T. Pfister, X. Li, G. Zhao, M. Pietikäinen, Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 868–875.

[30] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), IEEE, 2013, pp. 1–6.

[31] Y. Wang, J. See, R. C.-W. Phan, Y.-H. Oh, LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition, in: Asian conference on computer vision, Springer, 2014, pp. 525–537.

[32] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, M. Pietikäinen, Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods, IEEE transactions on affective computing 9 (4) (2017) 563–577.

[33] Y.-J. Liu, B.-J. Li, Y.-K. Lai, Sparse MDMO: Learning a discriminative feature for micro-expression recognition, IEEE Transactions on Affective Computing 12 (1) (2018) 254–261.

[34] S.-T. Liong, J. See, K. Wong, R. C.-W. Phan, Automatic micro-expression recognition from long video using a single spotted apex, in: Asian conference on computer vision, Springer, 2016, pp. 345–360.

[35] F. Zhang, T. Zhang, Q. Mao, C. Xu, Joint pose and expression modeling for facial expression recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3359–3368.

[36] B. Meden, Ž. Emeršič, V. Štruc, P. Peer, k-same-net: k-anonymity with generative deep neural networks for face deidentification, Entropy 20 (1) (2018) 60.

[37] M. Pernuš, V. Štruc, S. Dobrišek, High resolution face editing with masked gan latent code optimization, arXiv preprint arXiv:2103.11135.

[38] Ž. Emeršič, D. Sušanj, B. Meden, P. Peer, V. Štruc, ContexedNet: Context–aware ear detection in unconstrained settings, IEEE Access 9 (2021) 145175–145190.

[39] M. Ivanovska, V. Štruc, Y-GAN: Learning dual data representations for efficient anomaly detection, arXiv preprint arXiv:2109.14020.

[40] M. Peng, C. Wang, T. Chen, G. Liu, X. Fu, Dual temporal scale convolutional neural network for micro-expression recognition, Frontiers in psychology 8 (2017) 1745.

[41] N. Van Quang, J. Chun, T. Tokuyama, CapsuleNet for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–7.

[42] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, L.-K. Tan, OFF-ApexNet on micro-expression recognition system, Signal Processing: Image Communication 74 (2019) 129–139.

[43] J. Li, Y. Wang, J. See, W. Liu, Micro-expression recognition based on 3d flow convolutional neural network, Pattern Analysis and Applications 22 (4) (2019) 1331–1339.

[44] Z. Xia, X. Hong, X. Gao, X. Feng, G. Zhao, Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions, IEEE Transactions on Multimedia 22 (3) (2019) 626–640.

[45] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, L. Zhao, Recognizing spontaneous micro-expression using a three-stream convolutional neural network, IEEE Access 7 (2019) 184537–184551.

[46] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

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

[48] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2852–2861.

[49] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., Challenges in representation learning: A report on three machine learning contests, in: International conference on neural information processing, Springer, 2013, pp. 117–124.

[50] Y. Zhao, J. Xu, Compound micro-expression recognition system, in: 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE, 2020, pp. 728–733.

[51] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM transactions on graphics (TOG) 31 (4) (2012) 1–8.

[52] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, CASME II: An improved spontaneous micro-expression database and the baseline evaluation, PloS one 9 (1) (2014) e86041.

[53] A. K. Davison, C. Lansley, N. Costen, K. Tan, M. H. Yap, SAMM: A spontaneous micro-facial movement dataset, IEEE transactions on affective computing 9 (1) (2016) 116–129.

[54] G. Bradski, A. Kaehler, Learning OpenCV: Computer vision with the OpenCV library, " O'Reilly Media, Inc.", 2008.

[55] J. See, M. H. Yap, J. Li, X. Hong, S.-J. Wang, MEGC 2019-the second facial micro-expressions grand challenge, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[56] L. Zhou, Q. Mao, L. Xue, Dual-inception network for cross-database micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[57] J. Yu, C. Zhang, Y. Song, W. Cai, ICE-GAN: Identity-aware and capsule-enhanced gan for micro-expression recognition and synthesis, arXiv e-prints (2020) arXiv–2005.

[58] P. Gupta, MERASTC: Micro-expression recognition using effective feature encodings and 2d convolutional neural network, IEEE Transactions on Affective Computing 14 (2) (2021) 1431–1441.

[59] A. C. L. Ngo, R. C.-W. Phan, J. See, Spontaneous subtle expression recognition: Imbalanced databases and solutions, in: Asian conference on computer vision, Springer, 2014, pp. 33–48.

[60] F. Xu, J. Zhang, J. Z. Wang, Microexpression identification and categorization using a facial dynamics map, IEEE Transactions on Affective Computing 8 (2) (2017) 254–267.

[61] A. C. Le Ngo, Y.-H. Oh, R. C.-W. Phan, J. See, Eulerian emotion magnification for subtle expression recognition, in: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2016, pp. 1243–1247.

[62] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012) 1097–1105.

[63] H.-Q. Khor, J. See, R. C. W. Phan, W. Lin, Enriched long-term recurrent convolutional network for facial micro-expression recognition, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 667–674.

[64] H.-Q. Khor, J. See, S.-T. Liong, R. C. Phan, W. Lin, Dual-stream shallow networks for facial micro-expression recognition, in: 2019 IEEE international conference on image processing (ICIP), IEEE, 2019, pp. 36–40.

[65] S.-T. Liong, K. Wong, Micro-expression recognition using apex frame with phase information, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 534–537.

[66] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[67] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.