

This is the **accepted version** of the journal article:

Gong, Wenjuan; Zhang, Yue; Wang, Wei; [et al.]. «Meta-MMFNet : meta-learning based multi-model fusion network for micro-expression recognition». ACM transactions on multimedia computing, communications and applications, Vol. 20, issue 2 (February 2024), art. 39. DOI 10.1145/3539576

This version is available at <https://ddd.uab.cat/record/311822>

under the terms of the  **IN
COPYRIGHT** license

Meta-MMFNet: Meta-Learning Based Multi-Model Fusion Network for Micro-Expression Recognition

WENJUAN GONG and YUE ZHANG, China University of Petroleum (East China), China

WEI WANG, Beijing Institute for General Artificial Intelligence, China

PENG CHENG, Institute of High Performance Computing, A*STAR, Singapore

JORDI GONZÁLEZ, Computer Vision Center, Univ. Autònoma de Barcelona, Spain

Despite its wide applications in criminal investigations and clinical communications with patients suffering from autism, automatic micro-expression recognition remains a challenging problem because of the lack of training data and imbalanced classes problems. In this study, we proposed a meta-learning based multi-model fusion network (Meta-MMFNet) to solve the existing problems. The proposed method is based on the metric-based meta-learning pipeline, which is specifically designed for few-shot learning and is suitable for model-level fusion. The frame difference and optical flow features were fused, deep features were extracted from the fused feature, and finally in the meta-learning-based framework, weighted sum model fusion method was applied for micro-expression classification. Meta-MMFNet achieved better results than state-of-the-art methods on four datasets. The code is available at <https://github.com/wenjgong/meta-fusion-based-method>.

Additional Key Words and Phrases: Feature Fusion, Model Fusion, Meta-Learning, Micro-Expression Recognition

ACM Reference Format:

Wenjuan Gong, Yue Zhang, Wei Wang, Peng Cheng, and Jordi González. 2022. Meta-MMFNet: Meta-Learning Based Multi-Model Fusion Network for Micro-Expression Recognition. 1, 1 (March 2022), 20 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

1 INTRODUCTION

Facial expressions reveal emotions and are important non-verbal communication cues. Micro-expressions are unconscious momentary facial expressions that are difficult to recognize. They correlate with emotions accurately and reveal real emotions even when they are intentionally concealed [5]. Micro-expression recognitions are widely applicable in teaching evaluations [43], business negotiations [28], interrogations, and other fields.

One of the challenges of micro-expression recognition is the lack of micro-expression data. Deep learning models are essentially data-driven, thus more training data usually produces better model generalization, which hinders their applications in fields where data is scarce [1] as in micro-expression recognition. As a result, current research into micro-expression recognition mainly focuses on feature extraction. Recently, optical flow features based methods [46], [38] had dramatically improved the performances.

Authors' addresses: Wenjuan Gong, wenjuangong@upc.edu.cn; Yue Zhang, z20070048@s.upc.edu.cn, China University of Petroleum (East China), No. 66 Changjiangxi Road, Huangdao District, Qingdao, China, 266580; Wei Wang, Beijing Institute for General Artificial Intelligence, No.2 YiHeYuan Road, Haidian District, Beijing, China, wangwei@bigai.ai; Peng Cheng, Institute of High Performance Computing, A*STAR, 1 Fusionopolis Way, #16-16 Connexis (North Tower), Singapore, cheng_peng@ihpc.a-star.edu.sg; Jordi González, Computer Vision Center, Univ. Autònoma de Barcelona, 1 Fusionopolis Way, #16-16 Connexis (North Tower), Barcelona, Spain, Jordi.Gonzalez@uab.cat.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

Meta-learning based methods were proposed to deal with the lack of data problem. They are also known as methods of “learning to learn” [41], which were inspired by the observation that human can learn knowledge from a small amount of samples and generalize it to new data. A deep learning domain corresponds to a specific task, e.g., a cat-dog classification problem, whereas the meta-learning domain is composed of various tasks. Knowledge gained over training tasks enables quick adaptations to test tasks. Meta-learning has been successfully applied in few-shot image recognition [6] [37], reinforcement learning, neural architecture search (NAS) [22] [35], unsupervised learning [27], etc.

Meta-learning based methods are categorized into three classes: metric-based, optimization-based, and model-based methods. The core idea of metric-based meta-learning is similar to those of the nearest neighbor and kernel density estimation methods, which solves the problem by learning an effective distance measure. Snell et al. [37] proposed prototypical networks, which projected the support set into a feature space using clustering, calculated the average feature of each class, and classified the query set using a nearest centroid method. A recent study by Chen et al. [3] provided a metric-based meta-baseline, which pre-trained classifiers using all base classes, calculated the average feature of the support set, and classified the query samples using cosine distances. Optimization-based meta-learning methods extract meta knowledge to improve optimization performance. One representative model is the model-agnostic meta-learning (MAML) [7], which trained the initial model parameters so that it had maximal performance on a new task. Model-based meta-learning methods employ recurrent neural networks with explicit or implicit memories. One noteworthy study is the memory-augmented neural network based on neural Turing machines [36], which used an external memory storage to encode information explicitly and combined it with the long-term memory of the neural network.

In this study, we employed metric-based meta-learning because it provides a natural solution for information fusion. Information fusion is a classical machine learning method that fully exploits the input data [52] [29]. Generally, various forms of information, e.g., extracted features, are partially different and overlapping. These representations are usually fused in three ways: early fusion (a.k.a. feature fusion), deep fusion (fusion using deep neural networks), and late fusion (a.k.a. model fusion). Early fusion requires data alignment for feature addition, and deep fusion requires architecture design in the feature layer of deep neural networks [2]. In early fusion, features are concatenated [10] or added together [24] to form a final feature representation, whereas late fusion operates in a later stage, e.g., the recognition stage.

We focus on meta-learning fusion methods and explore the effect of various fusion methods for micro-expression recognition. Various combination of features were evaluated. Furthermore, we argue that the pre-trained model using macro-expressions and micro-expressions carry distinct information. Therefore, the prior knowledge of the micro-expression and macro-expression is fused in the meta-learning framework. Metric-based meta-learning provides a natural solution for model-level fusion, thus we utilize it as the basic architecture in this study. The contributions of this study are as the followings.

- (1) We proposed a novel meta-learning based method to solve the micro-expression recognition problem. Experiments showed that the proposed method was effective with a limited number of micro-expression data available.
- (2) We carried out feature and model fusions in the metric-based meta-learning framework and achieved better results than using a single feature or model.
- (3) We evaluated the proposed Meta-MMFNet on three publicly available micro-expression datasets and one composite dataset, and the proposed method achieved comparable performance to the state-of-the-art methods.

2 RELATED WORK

Existing micro-expression recognition methods are divided into two categories: traditional methods and deep learning based methods. Traditional methods generally extract features from the input images and then use the extracted information to optimize the parameters of the classic machine learning classifier. Deep learning method, on the other hand, use deep neural network models to extract features and classify the input images.

2.1 Traditional Methods based Micro-expression Recognition

Local binary pattern (LBP) [31] and its variants are widely used appearance features for face recognition. Zhao et al. [51] proposed a LBP on three orthogonal plane (LBP-TOP) feature to extract textures from input images. To reduce the redundant information in the LBP-TOP, Wang et al. [42] proposed a LBP with six intersection points (LBP-SIP) feature. The feature was generated by the intersection lines of three orthogonal planes to describe the dynamic facial texture information.

Huang et al. [12] not only extracted LBP features as the sign-based difference between a center pixel and its neighbor pixels, but also added direction and amplitude information, and the three components are combined to form a final representation. Niu et al. [30] proposed a local two-order gradient pattern (LTOGP) feature, which used eight masks to extract eight neighbor pixels' two-order gradients.

In order to capture facial motions, optical flow feature [39] [50] was also introduced. Traditional micro-expression recognition algorithms either operate over the entire micro-expression video sequence or a fraction of it. However, this might cause redundancy due to high frequency videos captured using high-speed cameras. Liong et al. [21] proposed that a video sequence could be denoted by its onset and apex frames, and proposed a bi-weighted oriented optical flow (BI-WOOF) feature, in which the computed optical flow histogram was weighted twice to highlight facial motions. Lu et al. [25] also used the onset and apex frames as inputs, and proposed a fusion of motion boundary histograms (FMBH) feature, which combined gradient vector fields of different components of the optical flow feature.

2.2 Deep Learning based Micro-expression Recognition Methods

Deep learning has boosted algorithm performances on many research topics, including micro-expression recognition. Commonly used deep learning models for micro-expression recognition includes convolutional neural networks (CNNs), long- and short-term memory networks (LSTMs), etc. For example, Kim et al. [14] and Peng et al. [32] used CNNs to encode spatial information, and LSTMs for temporal information.

In deep learning, data augmentation and transfer learning are two commonly used methods to deal with the lack of data problem. Takalkar et al. [40] carried out data augmentation by merging multiple datasets of micro-expression, and optimized CNN model parameters using the merged dataset. Alternatively, Peng et al. [33] performed transfer learning to solve this problem. They pre-trained the deep learning model using large-scale macro expression datasets, and fine-tuned the model parameters using micro-expression data.

To encode subtle facial movements, motion features are further introduced. Khor et al. [13] used optical flow and optical strain features as inputs which contain dynamic facial information. Based on the pre-trained Resnet18, Liu et al. [23] used adversarial training to learn domain invariant features from optical flow features extracted after applying Eulerian video magnification technology (EVM) [44]. Quang et al. [34] employed both transfer learning and data enhancement techniques to reduce the risk of overfitting during model training. Gan et al. [8] calculated the horizontal and vertical components of the optical flow feature between the onset and the apex frames, designed a dual-stream

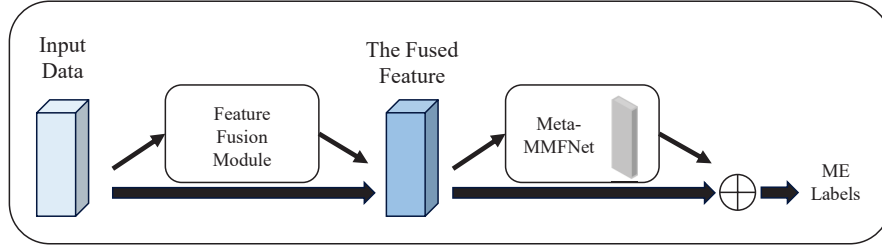


Fig. 1. Overall Structure of the Proposed Meta-MMFNet (Meta-learning based Multi-model Fusion Network) Method

network to extract the information for the two components. Outputs from the two streams are fused for further classification.

3 META-MMFNET FOR MICRO-EXPRESSION RECOGNITION

We consider micro-expression recognition as a few-shot classification problem and use a metric-based meta-learning based framework to solve it. We follow a standard N -way K -shot classification task definition, which is usually divided into two stages. The training stage is responsible for learning from the support set S , and the test stage involves predicting micro-expression labels for the query set Q . N denotes that the support set are from N different classes, and K is the number of labelled training samples in each class of a task. The query set data samples are also drawn from these N categories, and the goal of a N -way K -shot classification task is to classify unlabelled samples in the query set as one of the N categories.

Using this protocol, we experimented with various combinations of the optical flow and frame difference features calculated from the onset and apex frames. The fused features were input into the deep feature extraction model. We introduced prior knowledge from micro-expression and macro-expression data using pre-trained deep learning models, and studied various methods for fusing them in the metric-based meta-learning framework. Figure 1 illustrates the overall structure of this study.

We conducted experiments on three public micro-expression datasets: the Spontaneous Micro-expression Corpus (SMIC) Dataset, the Chinese Academy of Sciences Micro-expression (CASME) Dataset, and the CASME II Dataset, and achieved state-of-the-art performances. Furthermore, we synthesized the three datasets into a composite dataset and evaluated the proposed method on the composite dataset.

3.1 Data Preparation

First, we used Active Shape Model (ASM) [4] to detect and crop the face area. Faces were then aligned to a reference face based on extracted key points, as shown in Figure 2. Motion features extracted from aligned faces are intrinsically more effective descriptors. In addition, we carried out data augmentation to deal with the unbalanced class problem. Images from classes that contain fewer data samples were flipped horizontally. For example, for the CASME dataset, all categories except “tense” were flipped horizontally, and for the CASME II dataset, all data samples were flipped horizontally except those from the “others” category. Following the standard settings for evaluating few-shot learning approaches, micro-expression recognition problems on the CASME and CASME II datasets are formulated as 4-way 5-shot classification tasks, and that on the SMIC dataset as a 3-way 5-shot task.

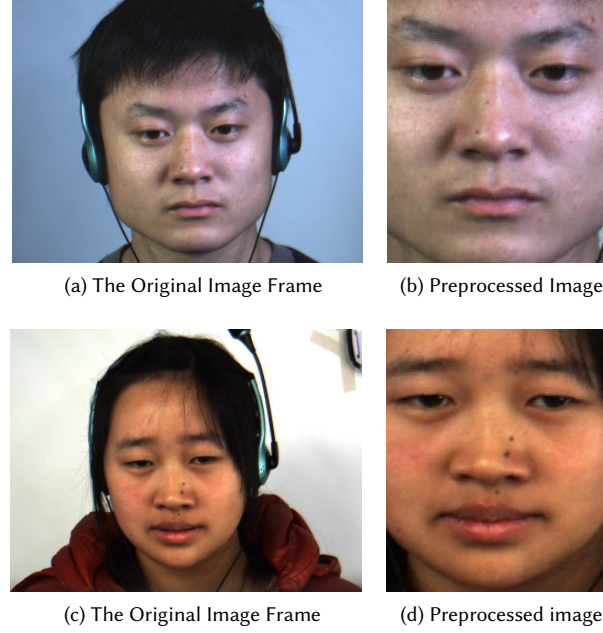


Fig. 2. Preprocessing of Image Frames

3.2 Feature Fusion Method

In this study, we consider two commonly used features for micro-expression recognition: the frame difference [19] and optical flow [16] features. These features capture subtle muscle movements crucial for micro-expression recognition. The frame difference feature computes per-pixel intensity variations between two frames, whereas the optical flow extracts the pixel motion between them.

Previous studies showed that features extracted between the onset and apex frames of video sequences were most effective for micro-expression recognition [20]. Therefore, we extracted the optical flow and frame difference features between these frames. The frame difference was computed on RGB channels, respectively. For the optical flow, we used the Gunnar Farneback algorithm to extract dense features. The horizontal and vertical components of the optical flow were further utilized to compute the motion magnitude. These three components were then normalized, respectively. The three normalized components constituted the three channels of the optical flow feature representation.

We evaluated three feature settings: one single feature of frame difference or optical flow, and a fused representation of two features concatenated together. Experiments showed that the fused feature achieved the best performance. It outperformed the frame difference with a margin of 14.64%, and the optical flow with a margin of 10.98% on the SMIC dataset. Thus, we used the fused feature as the final feature representation.

The optical flow and frame difference features are intrinsically complementary. The optical flow includes directions and magnitudes, whereas the frame difference computes the intensity variations per-pixel. If we examine features extracted from exemplary images of the “surprise” class in the three evaluated datasets, as shown in Fig. 3, we can observe that all three examples involve brow movements. The ground truth AU (action unit) annotations support this observation: the video sequences of the “surprise” class are characteristic for AU1 and AU2, both related with brow

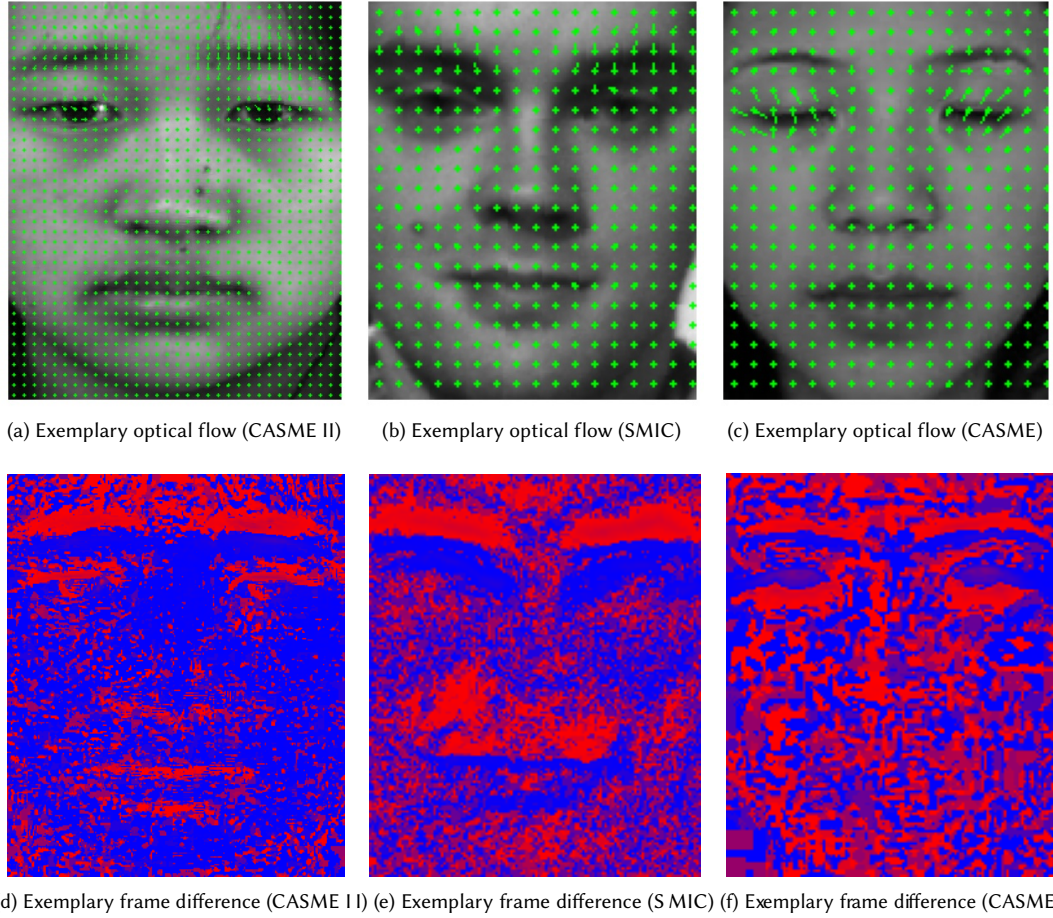


Fig. 3. Exemplary Optical Flow (On The First Row) From Three Datasets And Their Corresponding Frame Difference (On The Second Row). For Frame Difference, Red Colour Denotes Big Frame Difference Value and Blue Colour Denotes Small Value.

movements. We also observe that the when there are relatively fewer moving pixels (most of the pixels in the frame difference are marked with blue), frame difference feature prevails. For example, the frame difference outperformed the optical flow feature in micro-expression recognition with a margin of 7.14% for the CASME II dataset. On the contrary, when there are strong responses in both features as in the case of the CASME dataset, two features performed comparable and the fused feature outperformed with a margin of 5.56%.

3.3 Meta-Learning Based Multi-Model Fusion Network

The fused features were further processed by Resnet18 based feature extraction models. The extracted deep features were then fed into metric-based meta-learning pipelines for recognition. The metric-based pipeline computed support set centroid for every class and classified the query set using a nearest neighbour method. We employed the commonly

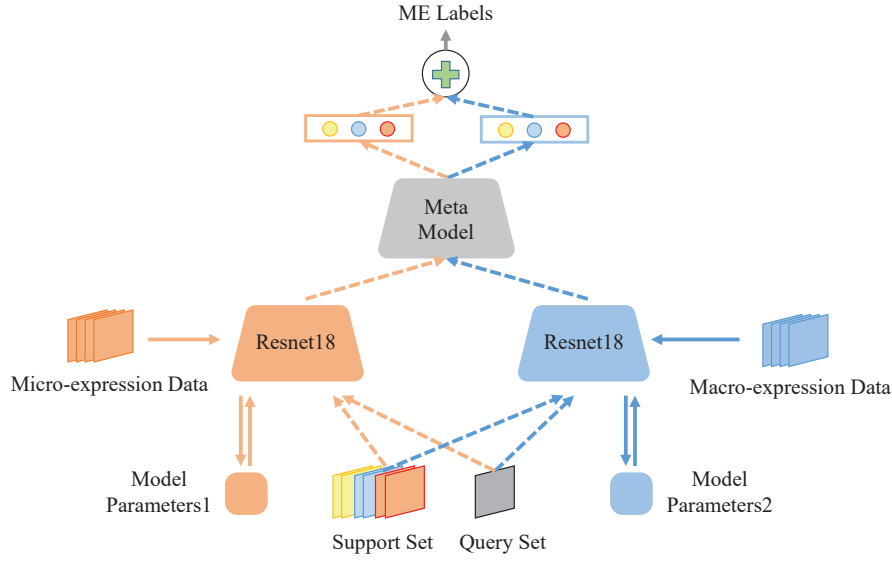


Fig. 4. The Network Structure of the Proposed Meta-MMFNet (Meta-learning based Multi-model Fusion Network) Method

used cosine similarity distance measure in the pipeline. We studied the fusion of micro-expression and macro-expression priors in the metric-based pipeline and proposed a meta-model-fusion method.

3.3.1 Network Structure. The Meta-MMFNet method consists of three modules: the deep feature extraction module, the metric-based distance computation module, and the model fusion module.

Related studies utilized prior knowledge from other domains through pre-trained deep neural networks. Based on this, we designed the deep feature extraction module to incorporate prior knowledge from micro- and macro-expressions. Then, a standard metric-based meta-learning pipeline was applied to calculate the mean feature for every class, and the cosine distances for every data sample of the query set. The last module aimed to fuse the incorporated micro- and macro-priors. We evaluated two fusion settings: concatenating deep features extracted using the Resnet18 based models, and weighted sum of cosine distances calculated using the metric-based pipelines. Experiments showed that the weighted sum of model fusion method outperformed the feature-level fusion method, on average, by 5.16%. Therefore, we chose the weighted sum model fusion method as the final solution. The network architecture of the proposed method is illustrated in Fig. 4. It consists of two data processing streams. The micro-stream encodes the micro-expression prior, and the macro-stream encodes the macro-expression prior. These two streams were fused using a model-fusion method.

3.3.2 Micro- and Macro-Priors. The deep feature extraction module used Resnet18 as its backbone. Resnet18 based models were optimized using the micro- and macro-expression data, respectively, in the pre-training phase to incorporate priors from these two domains. The fully connected layers for expression classification were then removed, and the remaining layers were utilized as deep feature extractors.

The deep feature vectors extracted using the above-mentioned feature extractors were sent to the two stream metric-based meta-learning pipelines, in which features from one class were averaged to compute a mean feature

vector:

$$w_c = \frac{1}{|S_c|} \sum_{x \in S_c} f_\theta(x), \quad (1)$$

where f_θ represents the feature extractor, $f_\theta(x)$ denotes the deep feature vector of data sample x , S_c is the sample cluster of class c in the support set S , and the result w_c is the centroid of class C .

We used cosine similarities to calculate the distance between the support set centroids and each sample of the query set. The distance measure was further processed by a softmax function to obtain the probability for further classification:

$$p_i^c = p(y = c|x_i) = \frac{\exp(< f_\theta(x_i), w_c >)}{\sum_{c'} \exp(< f_\theta(x_i), w_{c'} >)}, \quad (2)$$

where $< f_\theta(x), w_c >$ represents the cosine distance between the encoded query set feature $f_\theta(x)$ and centroid vector w_c .

The loss function was defined as the cross-entropy loss between the predicted distribution of the query set and the ground truth distribution:

$$Loss = - \sum_{i=1}^n Y_i \log P_i, \quad (3)$$

where $P_i = [p_i^1, p_i^2, \dots, p_i^C]$ is the predicted distribution, C is the total number of micro-expression categories, and $Y_i = [y_i^1, y_i^2, \dots, y_i^C]$ denotes the ground truth distribution of the i -th data sample. We used this loss to update micro- and macro-streams, respectively. Data stream incorporated with micro-prior (connected with orange dashed lines) and data stream with macro-prior (connected with blue dashed lines) were processed separately in two data streams, as illustrated in Fig. 4.

In the ablation study of the experiment section, we compared the performances of the micro- and macro-prior processing data streams. Experiments showed that the two data streams could be combined to enhance the overall classification performance. For example, for the “disgust” category of the CASME dataset, the prediction accuracy of the micro-model was 7.5% higher than that of the macro-model, and the prediction accuracy of the fused model was 1.79% higher than that of the micro-model. For the “happiness” category of the CASME II dataset, the prediction accuracy of macro-model was 12.5% higher than that of the micro-model and the prediction accuracy of fused model was 3.34% higher than that of the macro-model.

3.3.3 Weighted Sum Model Fusion. The two data streams were subsequently combined to enhance micro-expression recognition performance. The calculated distance of the two streams were added using a weighted sum:

$$d_{sum}^i = d_{micro}^i + \alpha d_{macro}^i, i \in \{1, 2, \dots, c\}, \quad (4)$$

where c is the total number of support set categories, α is the weight, and d_{micro}^i and d_{macro}^i denote the cosine similarity distances calculated in the micro- and macro-streams, respectively.

The weight α in Equation 4 was learned through experiments. Following the widely used micro-expression evaluation protocol, we applied leave-one-subject-out (LOSO) validation method. LOSO means to retain the sample data of one subject for verification and the sample data of all other subjects for training. The effectiveness of the method was then evaluated by iterating through all possible scenarios to calculate the overall accuracy. Once we selected the subject S_{te} for verification, another subject S_α was randomly selected to compute the weight α . All subjects, except the select two, formed the training set. Each value in the set $\alpha \in \{0.1, 0.2, \dots, 1\}$ was tested on the subject S_α , and the value with the

highest prediction accuracy was set as the final weight for predicting the label of the subject S_{te} . The actual weights used in the experiments are listed in Table 1.

Table 1. Weight Values in Evaluation

Dataset	CASME II	CASME	SMIC
Weight	1	0.4	0.1

The weighted distance value d_{sum}^i was then used to compute the nearest neighbour to predict the ME (micro-expression) class label.

4 EXPERIMENTS

4.1 Evaluation Datasets

Table 2. Compositions of the Evaluated Micro-expression Datasets

Datasets	Number Of Subjects	Number Of Samples	Frame Rate	Resolutions	AU and Apex Annotations
CASME	19	195	60	1280×720,640×480	✓
CASME II	26	255	200	640×480	✓
SMIC	16	164	100	640×480	×

We conducted experiments on three commonly used micro-expression datasets: the SMIC dataset, the CASME dataset, and the CASME II dataset. The compositions of these datasets are listed in Table 2.

The CASME dataset [48] consists of 195 video samples from 19 subjects. The frame rate of the micro-expression video sequences are 60fps, and the video frame resolutions are 1280×720 and 640×480 , respectively. All data samples are divided into eight expression categories: “happiness,” “surprise,” “disgust,” “sadness,” “tense,” “repression,” “fear,” and “contempt.” The data samples of these eight emotions are highly unbalanced.

The CASME II dataset [49] contains 255 video sequences, which were recorded from 26 subjects by a high-speed camera with a frame rate of 200fps. The video frame resolutions are 640×480 . This dataset contains seven expression categories: “happiness,” “others,” “disgust,” “repression,” “surprise,” “sadness,” and “fear.” Both CASME and CASME II datasets provide apex frame and AU annotations.

The SMIC dataset [17] consists of 164 video sequences, which were recorded with a frame rate of 100fps. The resolutions are 640×480 . Most of the subjects in this dataset are Asian. The dataset contains three expression categories: “positive,” “negative,” and “surprise.”

4.2 Experiment Settings

In the pre-training phase, we used a SGD optimizer with the momentum set 0.9, and the learning rate is 0.01. We trained 50 epochs with a batchsize of 128 on a GPU. The weight decays is 0.0005. In the meta-learning stage, we use the SGD optimizer with momentum of 0.9, batch size of 4 and fixed learning rate of 0.05, that is, each training batch contains four few-shot tasks to calculate the average loss. At the same time, during the training, for the three micro-expression data sets, we enhance the data by flipping part of the data horizontally, which alleviates the problem of data imbalance to a certain extent.

We solved the micro-expression recognition problem as few-shot classification tasks, namely a N-way K-shot classification problem. For the three evaluated datasets, K was all set as 5, and the value of N depended on the number of categories in the dataset. For example, for the SMIC data set, we performed 3-way 5-shot classification tasks, while for the CASME and CASME II datasets, we performed 4-way 5-shot classification tasks.

We adopted a leave-one-subject-out (LOSO) cross-validation evaluation method. Specifically, we sequentially left out one subject as test, and all the remaining subjects were used as training. This process is repeated n (the number of subjects in the micro-expression dataset) times, and the final accuracy was computed over all subjects.

The data cluster scale of each micro-expression category varies dramatically. Some data cluster contains very few data samples, thus we omitted these categories. For the CASME dataset, we studied expressions that consists of more than 10 samples. For the CASME II dataset, we considered expressions that consists of more than 10 samples except the category “others.” All categories from the SMIC dataset were included in the experiments.

The SMIC dataset does not provide apex frame annotations, thus we picked the middle frame between the onset and offset frames as the apex frame. Before feature extraction, we scaled all cropped sample images to the dimensions of 80×80 .

To evaluate the proposed method, we conducted single data set comparison experiment, model selection experiment, and composite database evaluation (CDE) experiment. We compared the performances based on prediction accuracies, i.e. the ratio of the correctly classified query samples to the whole query set.

4.3 Evaluation Metrics

Following the standard measurement, we evaluated the performance of the proposed method using prediction accuracy. Prediction accuracy is calculated as the proportion of correctly classified negative samples to the total number of negative samples, formulated as follows:

$$\text{acc} = \frac{N_{\text{correct}}}{N_{\text{total}}}, \quad (5)$$

where N_{correct} denoted the number of samples that are correctly predicted, and N_{total} represents the total number of samples.

In addition to prediction accuracy, we also computed and visualized confusion matrices. Confusion matrix also evaluates the performance of a classification method and describes the distinctions among classes. The row and column elements in a confusion matrix correspond to the ground-truth and predicted categories, respectively. And the diagonal elements denotes correctly classified samples. The confusion matrices provide an overview of classification performances over all categories.

4.4 Ablation Study

Table 3. Ablation Study

Models	Dataset		
	CASME	CASME II	SMIC
Micro-model	0.6815	0.7733	0.6220
Macro-model	0.6051	0.7733	0.5244
Meta-MMFNet	0.6959	0.8095	0.6313

In this section, we evaluated the proposed method quantitatively. We conducted ablation studies to observe the effectiveness of model fusion. We carried out three experiments on each of the three micro-expression datasets to evaluate performances of various model configurations.

- (1) micro-model (micro-prior processing data stream) experiment, in which the pre-trained model with micro-expression dataset were utilized as a feature extractor, followed by the metric-based meta-learning pipeline, and finally a micro-expression classification module.
- (2) Macro-model (macro-prior processing data stream) experiment, in which the pre-trained model with macro-expression dataset were utilized as a feature extractor, followed by the metric-based meta-learning pipeline, and finally a micro-expression classification module.
- (3) Meta-MMFNet experiment, in which micro- and macro-models were pre-trained using micro- and macro-expression datasets, respectively, and followed by the metric-based meta-learning pipeline, and finally a weighted sum fusion module for the nearest neighbour classification.

Table 3 evaluated the above-mentioned models for micro-expression recognition and compared average prediction accuracies over all categories. The micro-model performed better than the macro-module, and the fused model achieved the best performance among the three. In summary, the significant increase in accuracy clearly proves the effectiveness of the proposed model fusion strategy for micro-expression recognition.

4.5 Comparisons with the State-of-the-art Methods

Table 4. Performance Comparisons between the Proposed Meta-MMFNet and the State-of-the-art Methods on the CASME Dataset

Methods	CASME Dataset				
	Disgust	Surprise	Repression	Tense	Overall
STCLQP [12]	0.64	0.50	0.53	0.58	0.5731
LBP-SIP [42]	-	-	-	-	0.3684
FHOFO [9]	-	-	-	-	0.6599
MER-RCNN [45]	-	-	-	-	0.632
STLBP-RIP [11]	0.5682	0.6	0.4211	0.8136	0.5906
DiSTLBP-RIP [11]	0.7273	0.6	0.5263	0.6667	0.6433
LGCcon [18]	0.57	0.8	0.21	0.77	0.6082
3DFCNN [15]	-	-	-	-	0.5444
Our Macro-model	0.625	0.8889	0.4242	0.6061	0.6051
Our Micro-model	0.7	0.8889	0.5758	0.6667	0.6815
Our Meta-MMFNet	0.7179	0.9412	0.5357	0.6875	0.6959

Table 5. Performance Comparisons between the Proposed Meta-MMFNet and the State-of-the-art Methods on the CASME II Dataset

Methods	CASME II Dataset				
	Surprise	Repression	Happiness	Disgust	Overall
DTCM [26]	-	-	-	-	0.7206
Our Macro-model	0.8929	0.5926	0.6563	0.8571	0.7733
Our Micro-model	0.9643	0.5926	0.5313	0.8889	0.7733
Our Meta-MMFNet	0.8929	0.5926	0.6897	0.9206	0.8095

Table 6. Performance Comparisons between the Proposed Meta-MMFNet and the State-of-the-art Methods on the SMIC Dataset

Methods	SMIC Dataset			Overall
	Surprise	Positive	Negative	
LBP-SIP [42]	-	-	-	0.4212
FDM [47]	0.53	0.66	0.41	0.5488
MER-RCNN [45]	-	-	-	0.571
FHOFO [9]	-	-	-	0.5122
Hierarchical STLBP-IP [53]	-	-	-	0.6078
FR [52]	-	-	-	0.579
3DFCNN [15]	-	-	-	0.5549
Our Macro-model	0.6512	0.5686	0.4143	0.5244
Our Micro-model	0.7442	0.5294	0.6143	0.6220
Our Meta-MMFNet	0.7561	0.5600	0.6087	0.6313

We compared the proposed Meta-MMFNet method with the state-of-the-art methods, and listed the results in Table 4, Table 5, and Table 6. The comparison is performed under the same classification settings of the same database. We conducted rigorous experiments on three databases. It can be observed that the method based on meta-fusion shows good performance on all three data sets, and the experiments showed that our method achieved better performance than the 3DFCNN, LGCcon, and other methods.

From the tables, we observed that various methods achieved the highest accuracy for every category of the evaluated datasets. For example, the STLBP-RIP method achieved the highest prediction accuracy for the “tense” category of the CASME dataset, but its overall accuracy was lower than the proposed Meta-MMFNet method. Although the micro-model achieved the highest prediction accuracy for the “repression” category of the CASME dataset, the “surprise” category of the CASME II dataset, and the “negative” category of the SMIC dataset, the overall performance of the proposed Meta-MMFNet method outperformed all the compared methods.

The confusion matrices of the proposed method on the evaluated datasets are shown in Fig. 5. From the figures, we observed that the “surprise” and “disgust” categories were easy to distinguish from other categories. In contrast, the “repression,” “tense,” and “happiness” categories were relatively difficult to recognize. The “repression” and “tense” categories were the most easily confused categories.

Furthermore, we plotted the loss and accuracy value variations during the training process in Fig. 6. Figure 6 plotted the training curves of all subjects from the CASME II dataset. Subfigure (a) showed the training accuracy and loss variations using a model pre-trained with macro-expression dataset, and subfigure (b) showed the training accuracy and loss variations with micro-expression dataset. We observed that the proposed Meta-MMFNet model had higher generalization ability after pre-training. For each subject, it required only six epochs. In the 4th epoch, the loss and accuracy variations reached plateaus. Compared with the model pre-trained using macro-expression data, the overall training stage of the model pre-trained using the micro-expression data is more stable.

To provide an intuitive impression about the classification results, we visualized correctly and incorrectly predicted test examples for each category of the CASME (Fig. 7), CASME II (Fig. 8), and SMIC (Fig. 9) datasets. For each test instance, we showed three image frames, namely the onset, apex, and offset frames. Although the micro-expression action is subtle and hard to detect through visual comparison, we observed that existed facial motion variations between the correctly classified examples and misclassified examples. For metric-based methods, samples that were far away from the centroid of every category in the feature space may be excluded, thus those samples with special action units

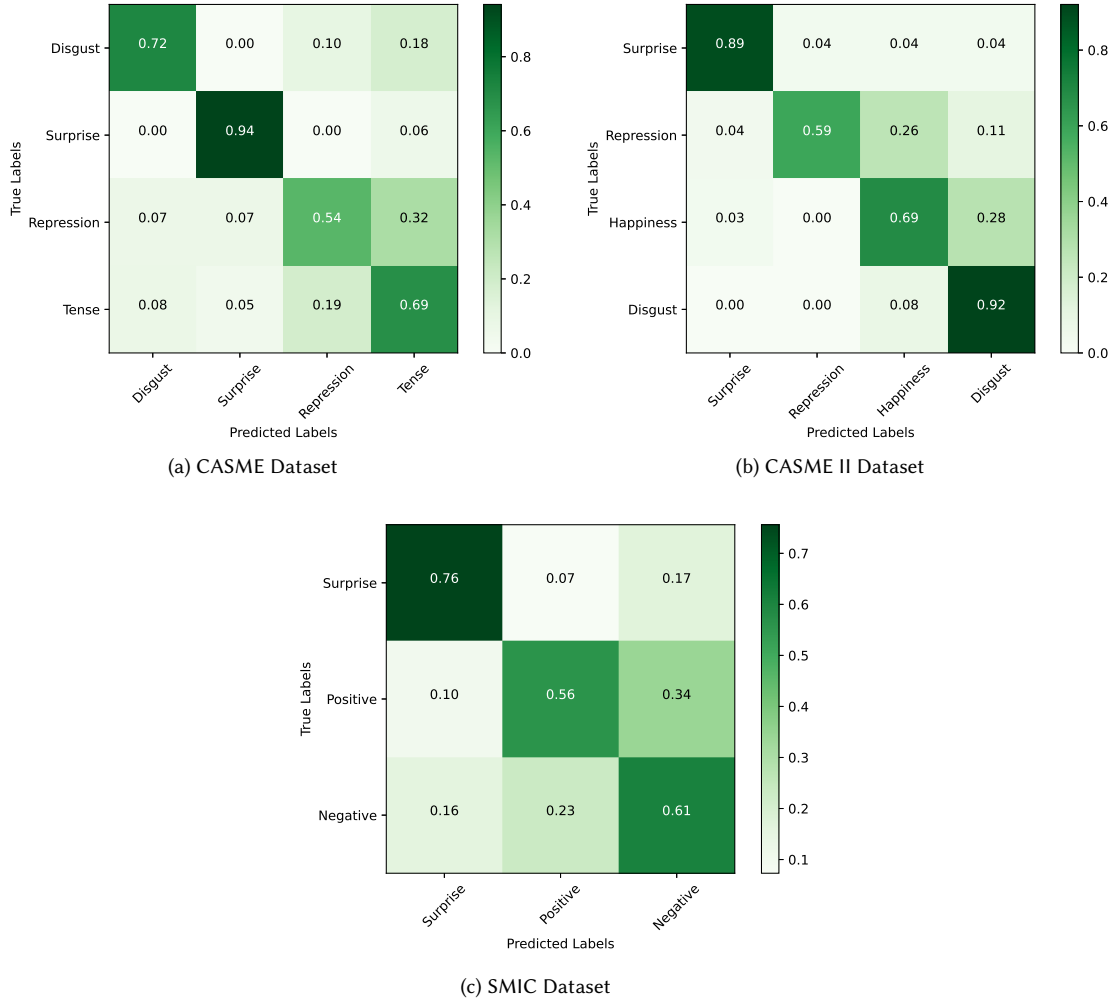


Fig. 5. Confusion Matrices of the Proposed Meta-MMFNet Method on the Evaluated Datasets

may not be correctly labelled. For example, for the CASME II dataset (Fig. 8), the ground-truth label of the sample in subfigure (b) is “surprise”, but it is misclassified as “repression”. Its ground-truth action unit is no. 25, AU no. 25 only occurred twice for all the “surprise” samples of the CASME II dataset. Thus, the feature vector of the test sample is far away from the class prototype, so it is difficult to get the correct classification.

4.6 Composite Database Evaluation

Furthermore, we synthesized data samples from the three micro-expression datasets to form one dataset and evaluated the proposed method on the composite dataset. Because the original labels of the three datasets were inconsistent, we

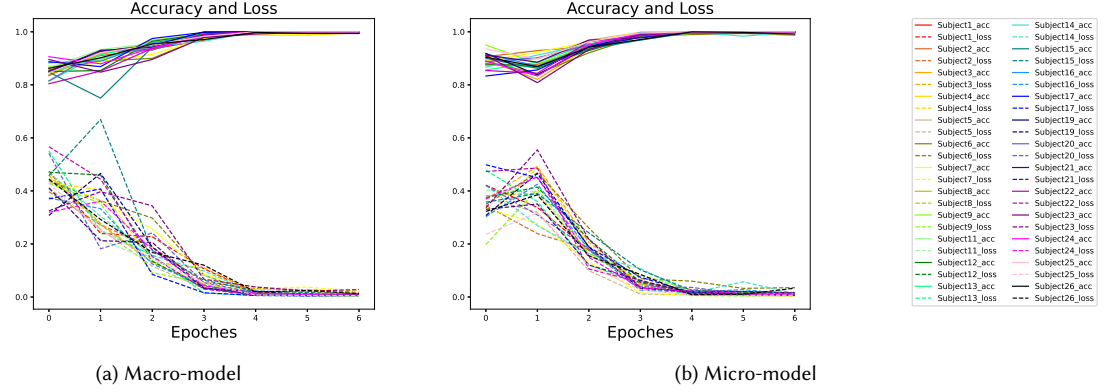


Fig. 6. Accuracy and Loss Variations of the proposed Meta-MMFNet model during Training on the CASME II dataset

Table 7. Composite Dataset Annotations

Datasets	Composite Database Evaluation (CDE) Dataset Categories		
	Surprise	Positive	Negative
CASME	Surprise	Happiness	Disgust, Repression, Tense, Sadness, Fear, Contempt
CASME II	Surprise	Happiness	Disgust, Repression, Fear, Sadness
SMIC	Surprise	Positive	Negative

relabelled the video sequences into three categories: “positive,” “negative,” and “surprise.” The relations between the original and composite dataset annotations are listed in Table 7.

Table 8. Experimental Results of the Composite Database Evaluation

Method	Surprise	Positive	Negative	Average Accuracy	Overall Accuracy
LBP-SIP [42]	-	-	-	0.3948	-
FHOFO [9]	-	-	-	0.5861	-
3DFCNN [15]	-	-	-	0.5497	-
Our Meta-MMFNet	-	-	-	0.7122	-
Our Meta-MMFNet_retrain	0.8736	0.5275	0.7855	-	0.7526

Composite Database Evaluation (CDE) evaluates the performance of an approach on a bigger and more complex dataset. Table 8 presents the experimental results of the proposed method on the composite dataset. For comparison, we computed the average prediction accuracies of other methods, e.g., the average prediction accuracy of the proposed Meta-MMFNet was the mean of the overall prediction accuracies on the three evaluated datasets. From the table, we can observe that by retraining on the CDE dataset, we enhanced the overall accuracy by 4.04% compared with the Meta-MMFNet method. Thus, even for few-shot learning, enriching the dataset by introducing more data samples and variations may enhance the model performance.

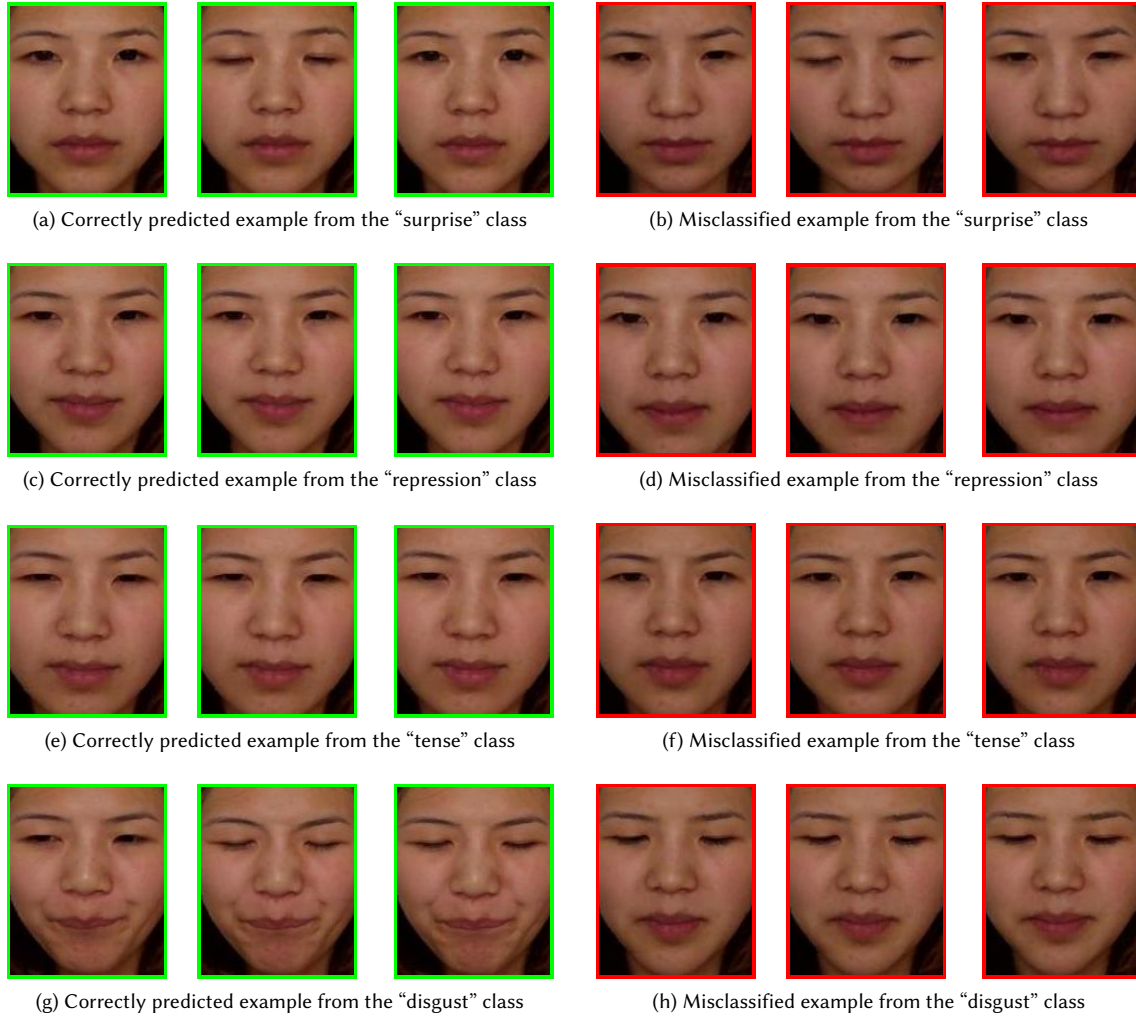


Fig. 7. Correctly and Incorrectly Classified Examples of the CASME Dataset

5 CONCLUSIONS AND FUTURE WORK

In this study, we investigated the application of meta-learning based multi-model fusion network(Meta-MMFNet) in the micro-expression recognition field. In contrast to existing deep-learning methods, we used meta-learning-based methods to process the fused feature of the frame difference and optical flow features. The fused features were further processed by a micro-model and a macro-model for deep feature extraction, and distance computations in a metric-based meta-learning pipelines, respectively. Finally, two distance measures were fused using a weighted sum and a micro-expression class label was predicted using a nearest neighbour method. The proposed Meta-MMFNet was evaluated on four databases: the CASME, CASME II, SMIC, and composite datasets. The experimental results indicated that the proposed Meta-MMFNet was accurate and efficient compared with the state-of-the-art methods. There is still

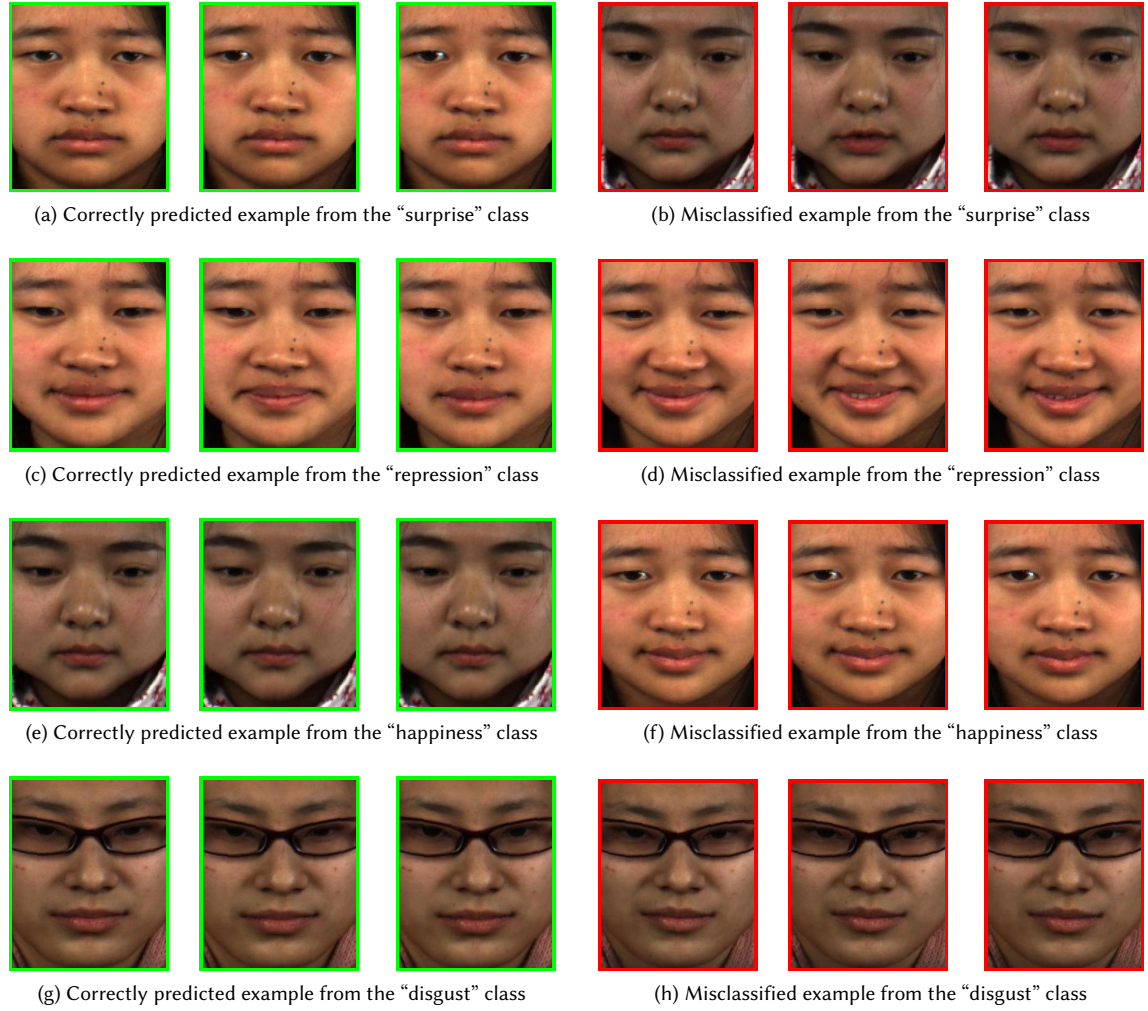


Fig. 8. Correctly and Incorrectly Classified Examples of the CASME II Dataset

much room for improvement for the proposed method. The standard procedure of computing optical flow features is separated with the recognition procedure so the input features are not tailored to fit the task and hinders further improvement of the proposed method.

In the future, we will explore other meta-learning based methods for solving the micro-expression recognition problem. First, we would like to explore more efficient feature extraction measures, because the optical flow feature, the standard feature used for micro-expression recognition, is hand crafted. In order to improve the performance, we would resort to extracting motion features using more efficient deep learning models, such as X3D networks, which extracts effective temporal features with a relatively small-scale network. Also, we would like to combine the proposed method with micro-expression detection method in the further work to obtain a holistic micro-expression analysis system. Currently, almost all micro-expression recognition methods are using additional annotations, such as ground

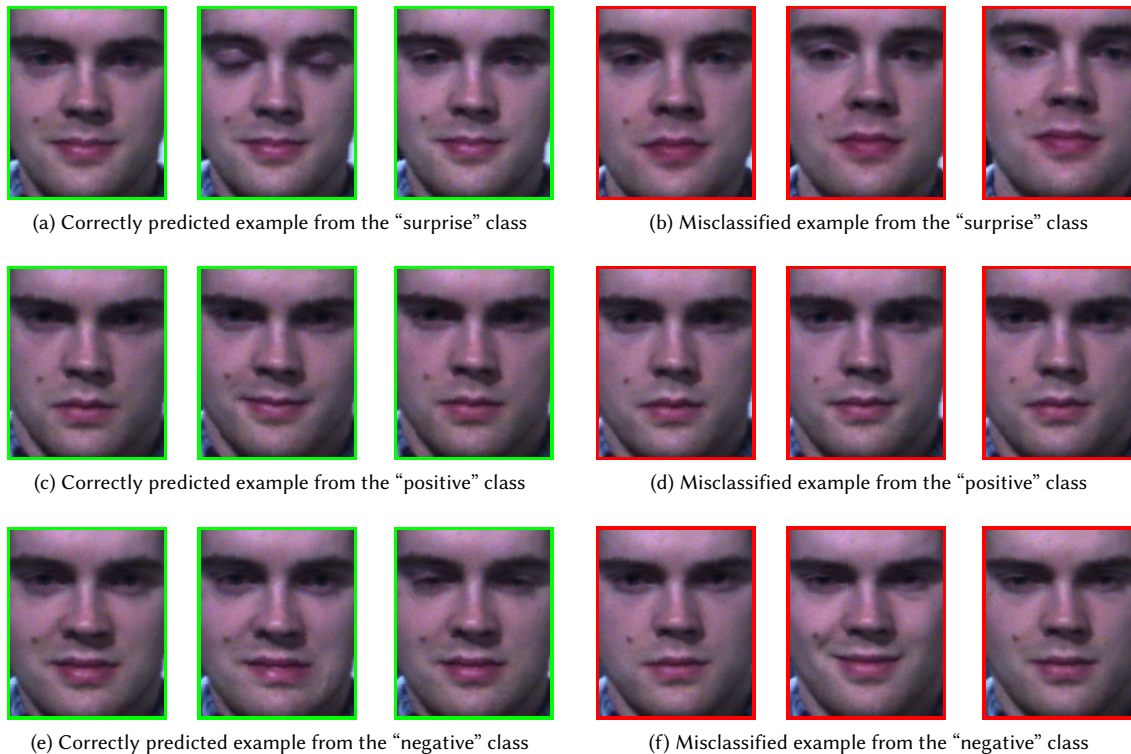


Fig. 9. Correctly and Incorrectly Classified Examples of the SMIC Dataset

truth apex frames and onset frames for computing input features. We believe through fusing multi-modal information from detection, we would improve applicability of the method by getting rid of these extra annotations.

ACKNOWLEDGMENTS

Jordi González acknowledges the support by the Spanish Ministry of Economy and Competitiveness (MINECO) and the European Regional Development Fund (ERDF) under Project PID2020-120311RB-I00/AEI/10.13039/501100011033.

REFERENCES

- [1] Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay S. Pande. 2016. Low Data Drug Discovery with One-shot Learning. *CoRR* abs/1611.03199 (2016). arXiv:1611.03199 <http://arxiv.org/abs/1611.03199>
- [2] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3D Object Detection Network for Autonomous Driving. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 6526–6534. <https://doi.org/10.1109/CVPR.2017.691>
- [3] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. 2020. A New Meta-Baseline for Few-Shot Learning. *CoRR* abs/2003.04390 (2020). arXiv:2003.04390 <https://arxiv.org/abs/2003.04390>
- [4] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. 1995. Active shape models-their training and application. *Computer vision and image understanding* 61, 1 (1995), 38–59.
- [5] P. Ekman. [n.d.]. *Lie Catching and Microexpressions*. The Philosophy of Deception. 118–133 pages.
- [6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research)*.

- Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, 1126–1135. <http://proceedings.mlr.press/v70/finn17a.html>
- [7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) (ICML '17). JMLR.org, 1126–1135.
- [8] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Tan Lit Ken. 2019. OFF-ApexNet on micro-expression recognition system. *Signal Process. Image Commun.* 74 (2019), 129–139. <https://doi.org/10.1016/j.image.2019.02.005>
- [9] S. L. Happy and Aurobinda Routray. 2019. Fuzzy Histogram of Optical Flow Orientations for Micro-Expression Recognition. *IEEE Trans. Affect. Comput.* 10, 3 (2019), 394–406. <https://doi.org/10.1109/TAFFC.2017.2723386>
- [10] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks. *CoRR abs/1608.06993* (2016). arXiv:1608.06993 <http://arxiv.org/abs/1608.06993>
- [11] Xiaohua Huang, Sujing Wang, Xin Liu, Guoying Zhao, Xiaoyi Feng, and Matti Pietikäinen. 2019. Discriminative Spatiotemporal Local Binary Pattern with Revisited Integral Projection for Spontaneous Facial Micro-Expression Recognition. *IEEE Trans. Affect. Comput.* 10, 1 (2019), 32–47. <https://doi.org/10.1109/TAFFC.2017.2713359>
- [12] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng, and Matti Pietikäinen. 2016. Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns. *Neurocomputing* 175 (2016), 564–578. <https://doi.org/10.1016/j.neucom.2015.10.096>
- [13] Huai-Qian Khor, John See, Raphael Chung-Wei Phan, and Weiyao Lin. 2018. Enriched Long-Term Recurrent Convolutional Network for Facial Micro-Expression Recognition. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*. IEEE Computer Society, 667–674. <https://doi.org/10.1109/FG.2018.00105>
- [14] Dae Hoe Kim, Wissam J. Baddar, and Yong Man Ro. 2016. Micro-Expression Recognition with Expression-State Constrained Spatio-Temporal Feature Representations. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, Alan Hanjalic, Cees Snoek, Marcel Worring, Dick C. A. Bulterman, Benoit Huet, Aisling Kelliher, Yiannis Kompatsiaris, and Jin Li (Eds.). ACM, 382–386. <https://doi.org/10.1145/2964284.2967247>
- [15] Jing Li, Yandan Wang, John See, and Wenbin Liu. 2019. Micro-expression recognition based on 3D flow convolutional neural network. *Pattern Anal. Appl.* 22, 4 (2019), 1331–1339. <https://doi.org/10.1007/s10044-018-0757-5>
- [16] Qiuyu Li, Shu Zhan, Liangfeng Xu, and Congzhong Wu. 2019. Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow. *Multim. Tools Appl.* 78, 20 (2019), 29307–29322. <https://doi.org/10.1007/s11042-018-6857-9>
- [17] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. 2013. A Spontaneous Micro-expression Database: Inducement, collection and baseline. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, Shanghai, China, 22-26 April, 2013*. IEEE Computer Society, 1–6. <https://doi.org/10.1109/FG.2013.6553717>
- [18] Yante Li, Xiaohua Huang, and Guoying Zhao. 2021. Joint Local and Global Information Learning With Single Apex Frame Detection for Micro-Expression Recognition. *IEEE Trans. Image Process.* 30 (2021), 249–263. <https://doi.org/10.1109/TIP.2020.3035042>
- [19] Chenhan Lin, Fei Long, JianMing Huang, and Jun Li. 2018. Micro-expression recognition based on spatiotemporal gabor filters. In *2018 Eighth International Conference on Information Science and Technology (ICIST)*. IEEE, 487–491.
- [20] Sze-Teng Liong, John See, Raphael Chung-Wei Phan, and KokSheik Wong. 2016. Less is More: Micro-expression Recognition from Video using Apex Frame. *CoRR abs/1606.01721* (2016). arXiv:1606.01721 <http://arxiv.org/abs/1606.01721>
- [21] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C.-W. Phan. 2018. Less is more: Micro-expression recognition from video using apex frame. *Signal Process. Image Commun.* 62 (2018), 82–92. <https://doi.org/10.1016/j.image.2017.11.006>
- [22] Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2019. DARTS: Differentiable Architecture Search. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. <https://openreview.net/forum?id=S1eYHoC5FX>
- [23] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. 2019. A Neural Micro-Expression Recognizer. In *14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019*. IEEE, 1–4. <https://doi.org/10.1109/FG.2019.8756583>
- [24] Yishu Liu, Yingbin Liu, and Liwang Ding. 2018. Scene Classification Based on Two-Stage Deep Feature Fusion. *IEEE Geosci. Remote. Sens. Lett.* 15, 2 (2018), 183–186. <https://doi.org/10.1109/LGRS.2017.2779469>
- [25] Hua Lu, Kidiyo Kpalma, and Joseph Ronsin. 2018. Motion descriptors for micro-expression recognition. *Signal Process. Image Commun.* 67 (2018), 108–117. <https://doi.org/10.1016/j.image.2018.05.014>
- [26] Zhaoyu Lu, Ziqi Luo, Huicheng Zheng, Jikai Chen, and Weihong Li. 2014. A Delaunay-Based Temporal Coding Model for Micro-expression Recognition. In *Computer Vision - ACCV 2014 Workshops - Singapore, Singapore, November 1-2, 2014, Revised Selected Papers, Part II (Lecture Notes in Computer Science)*, C. V. Jawahar and Shiguang Shan (Eds.), Vol. 9009. Springer, 698–711. https://doi.org/10.1007/978-3-319-16631-5_51
- [27] Luke Metz, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-Dickstein. 2019. Meta-Learning Update Rules for Unsupervised Representation Learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. <https://openreview.net/forum?id=HkNDsiC9KQ>
- [28] Michael W. Morris and Dacher Keltner. 2000. How Emotions Work: The Social Functions of Emotional Expression in Negotiations. *Research in Organizational Behavior* 22 (2000), 1–50. [https://doi.org/10.1016/S0191-3085\(00\)22002-9](https://doi.org/10.1016/S0191-3085(00)22002-9)
- [29] Nhi Thi Thu Nguyen, Duyen Thi Thu Nguyen, and Bao Thi Pham. 2021. Micro-expression recognition based on the fusion between optical flow and dynamic image. In *ICMLSC '21: 2021 The 5th International Conference on Machine Learning and Soft Computing, Da Nang, Vietnam, January 29-31, 2021*. ACM, 115–120. <https://doi.org/10.1145/3453800.3453821>

- [30] Mingyue Niu, Ya Li, Jianhua Tao, and Su-Jing Wang. 2018. Micro-Expression Recognition Based on Local Two-Order Gradient Pattern. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. 1–6. <https://doi.org/10.1109/ACIIAsia.2018.8470392>
- [31] Timo Ojala, Matti Pietikäinen, and David Harwood. 1994. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *12th IAPR International Conference on Pattern Recognition, Conference A: Computer Vision & Image Processing, ICPR 1994, Jerusalem, Israel, 9-13 October, 1994, Volume 1*. IEEE, 582–585. <https://doi.org/10.1109/ICPR.1994.576366>
- [32] Min Peng, Chongyang Wang, Tao Bi, Yu Shi, Xiangdong Zhou, and Tong Chen. 2019. A Novel Apex-Time Network for Cross-Dataset Micro-Expression Recognition. In *8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019, Cambridge, United Kingdom, September 3-6, 2019*. IEEE, 1–6. <https://doi.org/10.1109/ACII.2019.8925525>
- [33] Min Peng, Zhan Wu, Zhihao Zhang, and Tong Chen. 2018. From Macro to Micro Expression Recognition: Deep Learning on Small Datasets Using Transfer Learning. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*. IEEE Computer Society, 657–661. <https://doi.org/10.1109/FG.2018.00103>
- [34] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. 2019. CapsuleNet for Micro-Expression Recognition. In *14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019*. IEEE, 1–7. <https://doi.org/10.1109/FG.2019.8756544>
- [35] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. 2019. Regularized Evolution for Image Classifier Architecture Search. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 4780–4789. <https://doi.org/10.1609/aaai.v33i01.33014780>
- [36] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy P. Lillicrap. 2016. Meta-Learning with Memory-Augmented Neural Networks. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 (JMLR Workshop and Conference Proceedings)*, Maria-Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. JMLR.org, 1842–1850. <http://proceedings.mlr.press/v48/santoro16.html>
- [37] Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.), 4077–4087. <https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html>
- [38] Baolin Song, Ke Li, Yuan Zong, Jie Zhu, Wenming Zheng, Jingang Shi, and Li Zhao. 2019. Recognizing Spontaneous Micro-Expression Using a Three-Stream Convolutional Neural Network. *IEEE Access* 7 (2019), 184537–184551. <https://doi.org/10.1109/ACCESS.2019.2960629>
- [39] Deqing Sun, Stefan Roth, and Michael J. Black. 2014. A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles Behind Them. *Int. J. Comput. Vis.* 106, 2 (2014), 115–137. <https://doi.org/10.1007/s11263-013-0644-x>
- [40] Madhumita A. Takalkar and Min Xu. 2017. Image Based Facial Micro-Expression Recognition Using Deep Learning on Small Datasets. In *2017 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2017, Sydney, Australia, November 29 - December 1, 2017*. IEEE, 1–7. <https://doi.org/10.1109/DICTA.2017.8227443>
- [41] Sebastian Thrun and Lorian Y. Pratt. 1998. Learning to Learn: Introduction and Overview. In *Learning to Learn*, Sebastian Thrun and Lorian Y. Pratt (Eds.). Springer, 3–17. https://doi.org/10.1007/978-1-4615-5529-2_1
- [42] Yandan Wang, John See, Raphael C.-W. Phan, and Yee-Hui Oh. 2014. LBP with Six Intersection Points: Reducing Redundant Information in LBP-TOP for Micro-expression Recognition. In *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I (Lecture Notes in Computer Science)*, Daniel Cremers, Ian D. Reid, Hideo Saito, and Ming-Hsuan Yang (Eds.), Vol. 9003. Springer, 525–537. https://doi.org/10.1007/978-3-319-16865-4_34
- [43] Jacob Whitehill, Zewelani Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98. <https://doi.org/10.1109/TAFFC.2014.2316163>
- [44] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John V. Guttag, Frédéric Durand, and William T. Freeman. 2012. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.* 31, 4 (2012), 65:1–65:8. <https://doi.org/10.1145/2185520.2185561>
- [45] Zhaoqiang Xia, Xiaoyi Feng, Xiaopeng Hong, and Guoying Zhao. 2018. Spontaneous Facial Micro-expression Recognition via Deep Convolutional Network. In *Eighth International Conference on Image Processing Theory, Tools and Applications, IPTA 2018, Xi'an, China, November 7-10, 2018*. IEEE, 1–6. <https://doi.org/10.1109/IPTA.2018.8608119>
- [46] Zhaoqiang Xia, Xiaopeng Hong, Xingyu Gao, Xiaoyi Feng, and Guoying Zhao. 2020. Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-Expressions. *IEEE Trans. Multim.* 22, 3 (2020), 626–640. <https://doi.org/10.1109/TMM.2019.2931351>
- [47] Feng Xu, Junping Zhang, and James Z. Wang. 2017. Microexpression Identification and Categorization Using a Facial Dynamics Map. *IEEE Trans. Affect. Comput.* 8, 2 (2017), 254–267. <https://doi.org/10.1109/TAFFC.2016.2518162>
- [48] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Sujing Wang, and Xiaolan Fu. 2013. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, Shanghai, China, 22-26 April, 2013*. IEEE Computer Society, 1–7. <https://doi.org/10.1109/FG.2013.6553799>
- [49] W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, and X. Fu. 2014. CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. *Plos One* 9, 1 (2014), e86041.
- [50] Christopher Zach, Thomas Pock, and Horst Bischof. 2007. A Duality Based Approach for Realtime TV- L^1 Optical Flow. In *Pattern Recognition, 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007, Proceedings (Lecture Notes in Computer Science)*, Fred A. Hamprecht, Christoph

- 989 Schnörr, and Bernd Jähne (Eds.), Vol. 4713. Springer, 214–223. https://doi.org/10.1007/978-3-540-74936-3_22
- 990 [51] Guoying Zhao and Matti Pietikäinen. 2007. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions.
- 991 *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 6 (2007), 915–928. <https://doi.org/10.1109/TPAMI.2007.1110>
- 992 [52] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. 2021. Feature refinement: An expression-specific feature learning and
- 993 fusion method for micro-expression recognition. *CoRR* abs/2101.04838 (2021). arXiv:2101.04838 <https://arxiv.org/abs/2101.04838>
- 994 [53] Yuan Zong, Xiaohua Huang, Wenming Zheng, Zhen Cui, and Guoying Zhao. 2018. Learning From Hierarchical Spatiotemporal Descriptors for
- 995 Micro-Expression Recognition. *IEEE Trans. Multim.* 20, 11 (2018), 3160–3172. <https://doi.org/10.1109/TMM.2018.2820321>
- 996
- 997
- 998
- 999
- 1000
- 1001
- 1002
- 1003
- 1004
- 1005
- 1006
- 1007
- 1008
- 1009
- 1010
- 1011
- 1012
- 1013
- 1014
- 1015
- 1016
- 1017
- 1018
- 1019
- 1020
- 1021
- 1022
- 1023
- 1024
- 1025
- 1026
- 1027
- 1028
- 1029
- 1030
- 1031
- 1032
- 1033
- 1034
- 1035
- 1036
- 1037
- 1038
- 1039
- 1040