Learning from Macro-expression: a Micro-expression Recognition Framework

Bin Xia[†1], Weikang Wang[†1], Shangfei Wang[*12] and Enhong Chen[3]
xiabin@mail.ustc.edu.cn,wkwang0916@outlook.com,{sfwang,cheneh}@ustc.edu.cn
[1]Key Lab of Computing and Communication Software of Anhui Province,
School of Computer Science and Technology, University of Science and Technology of China
[2]Anhui Robot Technology Standard Innovation Base
[3]Anhui Province Key Lab of Big Data Analysis and Application,
School of Computer Science and Technology, University of Science and Technology of China

ABSTRACT

As one of the most important forms of psychological behaviors, micro-expression can reveal the real emotion. However, the existing labeled micro-expression samples are limited to train a high performance micro-expression classifier. Since micro-expression and macro-expression share some similarities in facial muscle movements and texture changes, in this paper we propose a micro-expression recognition framework that leverages macro-expression samples as guidance. Specifically, we first introduce two Expression-Identity Disentangle Network, named MicroNet and MacroNet, as the feature extractor to disentangle expression-related features for micro and macro expression samples. Then MacroNet is fixed and used to guide the fine-tuning of MicroNet from both label and feature space. Adversarial learning strategy and triplet loss are added upon feature level between the MicroNet and MacroNet, so the MicroNet can efficiently capture the shared features of micro-expression and macro-expression samples. Loss inequality regularization is imposed to the label space to make the output of MicroNet converge to that of MicroNet. Comprehensive experiments on three public spontaneous micro-expression databases, i.e., SMIC, CASME2 and SAMM demonstrate the superiority of the proposed method.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

micro-expression recognition, macro-expression, adversarial learning

†Equal contribution.
∗Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413774

ACM Reference Format:
Bin Xia, Weikang Wang, shangfei Wang and Enhong Chen. 2020. Learning from Macro-expression: a Micro-expression Recognition Framework. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413774

1 INTRODUCTION

Micro-expressions are discovered by Ekman and Friesen in the process of examining filmed interview of a psychotic patient [3]. With films played in slow motion mode, they found that the patient was showing a very brief sad face between a long period false smile in order to hide her suicidal tendency. Compared to large-intensity and long duration characteristics of macro-expressions, micro-expressions are very brief and subtle facial expressions which normally occur when a person either deliberately or unconsciously conceals his or her genuine emotions [3, 4]. It always takes human beings lots of time to perceive and recognize them. Thus developing micro-expression recognition systems becomes necessary.

There are lots of efforts devoted to micro-expression recognition, which fall into two main kinds: handcraft feature methods [7–9, 15, 20, 22, 35] and deep feature methods [5, 12, 17, 19, 21, 30, 34, 40]. However, though easily implementing and embracing good geometric or spatiotemporal interpretations, handcraft features are not robust in the micro-expressions identification and classification, due to micro-expression's short duration and low intensity. As for deep networks, though powerful, they are limited by the scarcity of micro-expression databases. Only enough data can we use to implement efficient deep network with good generalization ability.

On the contrary, there are large amounts of macro-expression databases, each of which consists of vast labeled training samples. Macro-expression is voluntary facial expressions, and covers a large face area. Macro-expression is also characterized by high intensities, in terms of facial muscle movements and texture changes. By contrast, micro-expression is characterized by rapid facial movements and covers restricted facial area, it conveys hidden emotions that determine true human feelings and state-of mind. Although macro-expression has longer duration and higher intensity than micro-expression, these two expressions share some similarities in facial muscle movements and texture changes. Figure 1 shows a comparison between micro and macro expressions. We can obviously find that the surprise from micro-expression and from macro-expression both have raised eyebrows and opened eyes. And for

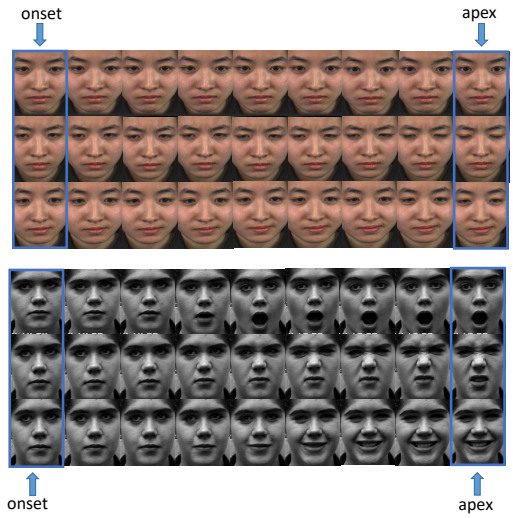

Figure 1: Here we give examples of micro-expression and macro-expression, where the above three rows are from CASME2 database and the below three rows are from CK+ database, both with surprise, disgust and happiness expression labels. Nine frames are chosen from each micro-expression or macro-expression video that change from onset to apex frame for each row.

the disgust, micro and macro-expression both show that the upper lip is slightly raised and the brows are slightly lowered, producing some wrinkling across the bridge of nose. For the happiness, micro and macro-expression both show that the lip corners are raised obliquiely in a slight smile, and there's also a deepening of the naso-labial fold that goes from outer corners of the nostrils down to the lip corners. Thus, how to take advantage of the macro-expression databsets for micro-expression recognition has become an important direction of the research.

In order to address problems mentioned above, we propose a micro-expression recognition framework that leverages macro-expression as guidance. Since subjects in macro-expression and micro-expression databases are different, Expression-Identity Disentangle Network (EIDNet) is introduced as feature extractor to disentangle expression-related features for expression samples. Specifically, we pretrain two EIDNets with micro and macro expression databases separately, named MicroNet and MacroNet. Then MacroNet is fixed and used to guide the fine-tuning of MicroNet from both label space and feature space, named Macro-to-Micro Network (MTMNet). By adversarial learning and triplet loss that added upon feature level between MicroNet and MacroNet, the MicroNet can learn shared representation from macro-expression samples. Furthermore, the loss inequality regularization is imposed in the label space to calibrate the output values of the MicroNet. Thus the proposed method can exploit patterns involved in the macro-expression samples to improve micro-expression recognition performance.

2 RELATED WORK

2.1 Micro-Expression Recognition

Micro-expression recognition methods can be categorized into two main kinds as handcrafted feature methods [7–9, 15, 20, 22, 35] and deep feature methods [5, 12, 17, 19, 21, 30, 34, 40].

Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOOF) and Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) based methods are the most prevailing handcraft feature methods. Li *et al.* [15] used feature difference for micro-expression spotting and adopted a variant of HOG for micro-expression recognition. For optical flow based methods, Liu *et al.* [22] proposed the Main Directional Mean Optical-flow (MDMO) method to describe micro-expressions and showed its superiority against LBP-TOP and HOOF features. Liong *et al.* [20] adopted a Bi-Weighted Oriented Optical Flow (Bi-WOOF) based feature extractor, while Happy *et al.* [7] proposed a fuzzy HOOF method, which ignored the subtle motion magnitudes and only take the motion direction into consideration. Le *et al.* [13] used LBP-TOP and sparsity constraints to learn significant temporal and spectral structures of micro-expressions. Huang *et al.* [8, 9] used LBP-TOP to extract the appearance and motion features on horizontal and vertical projections. Wang *et al.* [35] adopted a pruned LBP descriptor using six neighbors around every point.These methods greatly enhance the performance of micro-expression recognition, but due to micro-expression's short duration and low intensity, these methods are not robust.

Many researchers turn to the deep learning method as the micro expression databases gradually developed. Kim *et al.* [12] proposed two-phases method that used Convolutional Neural Network (CNN) to extract expression embeddings and adopted long short term memory recurrent neural network (LSTM-RNN) to do recognition. Liong *et al.* [5] proposed a CNN framework with optimal flow between the apex frame and onset frame as input features. Li *et al.* [17] used only apex frame passing through a deep neutral network to get features. Since standard CNNs are limited by their weakness in representing part-global relation, Nguyen *et al.* [34] adopted the newly proposed framework CapsuleNet [30] to recognize micro-expressions. Peng *et al.* [28] explored the underlying joint formulation for Motion MAGnification (MAG) and Time Interpolation Model (TIM) and proposed a consolidated framework for revealing the spatial-temporal information in micro-expression clips. Khor *et al.* [11] presented a Dual-Stream Shallow Network (DSSN) which robustly learns deep micro-expression features by exploiting a pair of shallow CNNs with heterogeneous motion-based inputs. Ling *et al.* [40] proposed a two-stream two-block variant of Inception network to learn a robust feature representation from horizontal and vertical optimal flow features. Motivated by the observation that deep CNN architectures do not perform well under limited micro-expression data, Liong *et al.* [19] aimed to put up with a compressed deep architecture with optimal flow features as inputs. Nevertheless, these deep learning methods suffer from insufficient training samples.

Due to the lack of large-scale micro-expression datasets, there have been a few works embracing the ideas of using macro-expression images or action unit information to assist micro-expression recognition. Sun *et al.* [32] proposed a knowledge transfer technique that distilled and transferred multi-knowledge from action unit for micro-expression recognition. Sun *et al.* pretrained a teacher

network on action unit recognition, and transferred it to a student network by penalizing the difference between the features of teacher network and the features of student network. Since teacher and student network did different tasks, making their features similar forcibly would cause that student network lose some domain specific information. Peng *et al.* [27] adopted transfer learning protocols to train a micro-expression recognition network pretrained on macro-expression database. Since they did not take the gap between micro-expression and macro-expression images into account, transfer learning did not achieve the desired effect. Jia *et al.* [10] and Ben *et al.* [1] extracted LBP-TOP features from micro-expression images and LBP features from macro-expression images. Jia *et al.* used singular value decomposition to achieve macro-to-micro transformation model, while Ben *et al.* employed coupled metric learning algorithm to model the shared features between micro-expression and macro-expression samples. These two methods used different extractors for macro and micro expression images, therefore, they can't effectively encoder the common features of macro and micro expression. Liu *et al.* [21] used Expression Magnification and Reduction (EMR) to reduce the gap between micro and macro expression. This preprocessing caused micro and macro expression visually similar, but this can't guarantee the similarity of micro and macro expression features, thus directly training on a fusion of micro and macro expression database can't generate appropriate expression-related features. While, in this paper we use adversarial learning strategy and triplet loss to model the shared features of micro-expression and macro-expression images.

2.2 Feature Disentanglement

Feature disentanglement technique aims to disentangle different kinds of features from original inputs for specific uses, it boosts the model performance since more domain related features are provided. We focus on related disentangle works in expression recognition areas in this section. There are two main directions: 1) disentangle facial expression apart from pose or head motions. 2) disentangle facial expression apart from identity. Li *et al.* [18] proposed a self-supervised disentangle auto-encoder for distinguishing AU-related features from motion-related features. Tran *et al.* [33] proposed a disentangled representation learning-generative adversarial network (DR-GAN) for learning facial expression apart from pose variances. Using the feature representation produced by the multi-scale contractive convolutional network (CCNET), Rifai *et al.* [29] trained a Contractive Discriminative Analysis (CDA) feature extractor to learn a representation separating out the emotion-related factors from the others. We are among the first to introduce feature disentanglement technique in micro-expression recognition. We obtain expression embeddings apart from identity embeddings, and thus efficiently leverage macro-expression images.

Compared with related work, our contributions are two-folds: 1) We propose a well-designed deep learning framework for micro-expression recognition by leveraging macro-expression databases as guidance. 2) We use Expression-Subject Disentanglement Network (EIDNet) to disentangle expression-related features apart from subject-related features from micro-expression and macro-expression images, and thus make the assistance of macro-expression more effectively.

3 METHOD

Our goal is to train a deep network for micro-expression recognition task with macro-expression databases as guidance. As shown in Figure 2, the training process can be splited into two phases. Firstly, a micro Expression-Identity Disentangle Network (MicroNet) is pretrained on the micro-expression training set $\mathcal{D}_I = \{x_N^{(i)}, x_E^{(i)}, y^{(i)}\}_{i=1}^{M_1}$, where $x_N^{(i)}, x_E^{(i)}$ are neutral and expression images of the same video from a micro-expression database, $y^{(i)} \in \{0, 1, \cdots, L-1\}$ is the expression label and M_1 is the number of videos. Simultaneously, a macro Expression-Identity Disentangle Network (MacroNet) is pretrained on the macro-expression training set $\mathcal{D}_A = \{x_N^{(j)}, x_E^{j}, y^{(j)}\}_{j=1}^{M_2}$, where $x_N^{(j)}, x_E^{(j)}$ are neutral and expression images of the same video from a macro-expression database, $y^{(j)} \in \{0, 1, \cdots, L-1\}$, is the corresponding expression label. Secondly, pretrained MicroNet is fine tuned on the micro and macro expression mixed training set $\mathcal{D}_{train} = \{x_{anc}^{(k)}, x_{pos}^{(k)}, x_{neg}^{(k)}, y^{(k)}\}_{k=1}^{M}$ with MacroNet as assistance, where $x_{anc}^{(k)}, x_{neg}^{(k)}$ are micro-expression images with different labels, while $x_{pos}^{(k)}$ is macro-expression image with the same label as $x_{anc}^{(k)}$ and $y^{(k)}$ is the corresponding expression label of $x_{anc}^{(k)}$. Finally, we can use the fine-tuned MicroNet to make a better prediction.

3.1 The Expression-Identity Disentangle Network

The whole structure of EIDNet is shown on the left of figure 2. The inputs of each EIDNet are pairs of images derived from the same video, consisting of a neutral facial image x_N and an expression facial image x_E. EIDNet mainly consists of four parts, named feature extraction, expression reconstruction, identity classification and expression classification. In feature extraction, EIDNet learns the features by respectively extracting the identity-related and expression-related features from two images through the two-branches encoder E. In expression reconstruction, decoder D_r integrates the identity features of neutral facial image and the expression features of expression facial image and uses them to reconstruct the expression facial image, ensuring that these two features are sufficient to represent the input expression image. Due to the differences from micro-expression and facial expression recognition problems, expression classifier D_e is added for the encoder to learn some domain specific information. We also introduce a identity classifier D_s and make the encoder E to extract expression-related feature that is easily recognized by the expression classifier D_e but difficult for identity classifier D_s to recognize.

3.1.1 Feature Extraction. We adopt a two-branches encoder to get expression-related features and identity-related features. EIDNet encodes the neutral facial image x_N and the expression facial image x_E by the encoder and gets their embeddings, $[f_N^e, f_N^s]$ and $[f_E^e, f_E^s]$, separately. To our understanding, neutral facial image only embraces identity information while expression facial image embraces both identity and expression information. Since these two images are from the same identity, their identity-related features, i.e., f_N^s and f_E^s should be similar, as shown in Eq (1):

$$\mathcal{L}_{sim} = \|f_E^s - f_N^s\|_2 \tag{1}$$

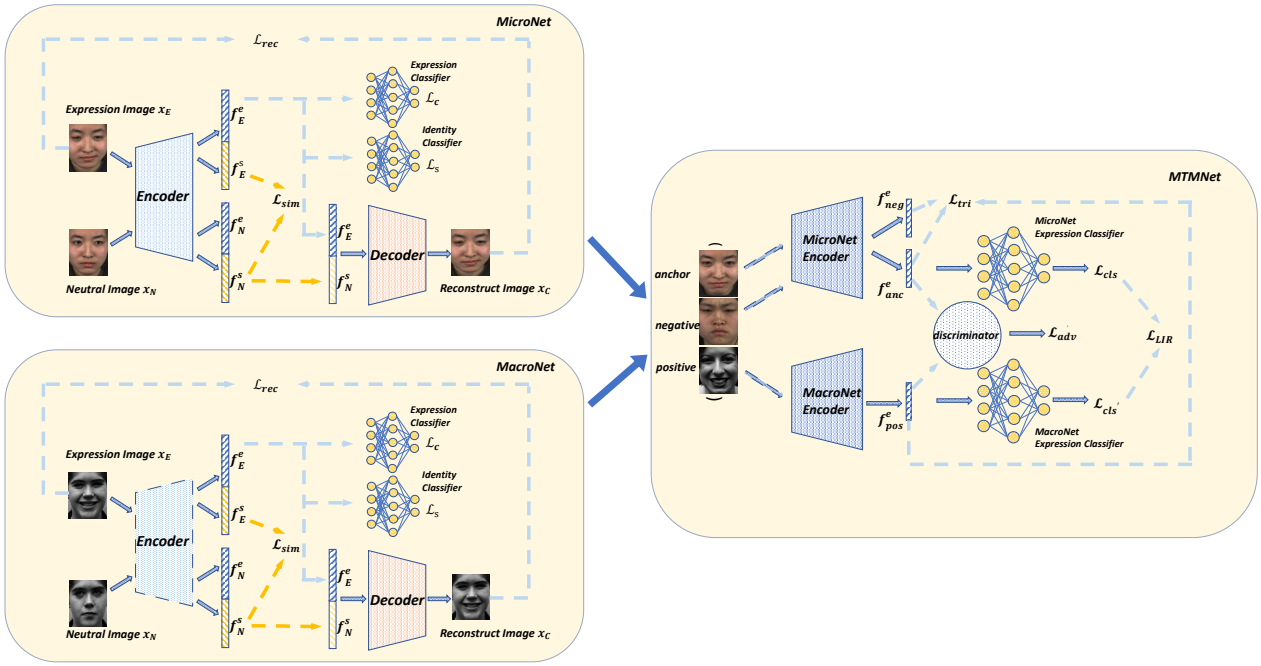

Figure 2: The framework of our micro-expression recognition model. First we pretrain two EIDNets with micro-expression and macro-expression databases separately, named MicroNet and MacroNet. Secondly, MacroNet is fixed and used to guide the fine-tuning of MicroNet from both label and feature space, named MTMNet.

3.1.2 Expression Reconstruction. In our method, we introduce a reconstruction loss. We believe that if we can reconstruct the original expression facial image x_E from concatenating f_N^e and f_E^s, the expression-related features f_E^e actually conveys enough expression information of x_E. We introduce a decoder D_r in the EIDNet to accomplish reconstruction. The expression reconstruction loss is:

$$\mathcal{L}_{rec} = \|x_E - D_r([f_E^e, f_N^s])\|_2 \tag{2}$$

3.1.3 Identity Classification. Since we want the extracted expression-related features contain no information of identity, expression-related features would have poor performance on identity classification task. It means the learned expression-related features are not discriminative for identity classifier D_s. Adversarial learning is introduced to accomplish this. Specifically, E and D_s play an adversarial game where E tries to minimize the divergence of feature distributions for different subject so that D_s has difficulty classifying the subject of expression-related features f_E^e. Thus we use the cross-entropy between predicted class distribution $D_s(f_E^e)$ and ground truth subject label s as the adversarial training objective:

$$\min_{D_s} \max_{E} \mathcal{L}_s = \min_{D_s} \max_{E} -\sum_{m=1}^{M} 1_{[s=m]} \log(D_s(f_E^e)) \tag{3}$$

3.1.4 Expression Classification. Since our goal is to produce pretrained feature extractor network for Macro-to-Micro Network, adopting expression classifier D_e can help to force features more suitable for their own domains, i.e., micro-expression and macro-expression recognition tasks. We conduct expression classification in EIDNet by adding expression classifier D_e after the encoder branch producing expression-related features. The Cross Entropy loss is used:

$$\mathcal{L}_c = -\sum_{l=1}^{L} 1_{[y=l]} \log(D_e(f_E^e)) \tag{4}$$

where y is the expression label of x_E, and $D_e(f_E^e)$ is predicted expression class distribution.

We train EIDNet to extract expression-related features by the losses described above, thus the overall loss of EIDNet is:

$$\mathcal{L}_{EID} = \min_{E, D_e} \max_{D_s} \mathcal{L}_c + \lambda_{1,1}\mathcal{L}_{rec} + \lambda_{1,2}\mathcal{L}_{sim} - \lambda_{1,3}\mathcal{L}_s \tag{5}$$

where $\lambda_{1,1}$, $\lambda_{1,2}$ and $\lambda_{1,3}$ are the hyperparameters controlling loss coefficients.

3.2 Macro-to-Micro Network

After pretraining MicroNet and MacroNet as described in the above section, in the Macro-to-Micro Network (MTMNet), we fix MacroNet and use it to guide the fine-tuning of MicroNet from both label and feature space, as shown on the right of Figure 2. Since all micro-expression images have corresponding macro-expression images with the same expression label, triplet term is adopted. We introduce adversarial learning to efficiently model the shared features of micro-expression and macro-expression. Expression classifier is used to control recognition accuracy of MicroNet and a new loss inequality regularization term is introduced to guide classification loss between MicroNet and MacroNet.

3.2.1 Guidance in Feature Space. MacroNet is fixed while MicroNet is further trained. Triplet inputs will be used, which consist of a micro-expression anchor x_{anc}, a same label macro-expression positive x_{pos} and a different label micro-expression negative x_{neg}. For every triplet, the anchor and negative will be passed through the MicroNet and the positive will be passed through the MacroNet. Three corresponding expression embeddings can be get: f^e_{anc}, f^e_{pos} and f^e_{neg}. The triplet loss is introduced at the feature level:

$$\mathcal{L}_{tri} = max\{\|f^e_{anc} - f^e_{pos}\|_2 - \|f^e_{anc} - f^e_{neg}\|_2 + m, 0\} \quad (6)$$

where m is the hyperparameter to guide the margin between these two distances.

An adversarial learning protocol is added between MicroNet and MacroNet. Since micro-expression anchor and macro-expression positive in one triplet have the same label, we hope that by adopting adversarial learning, their expression embeddings can show similar distributions. The fixed MacroNet offers expression embeddings of macro-expression images which are tagged as true labels; while MicroNet acts as the generator to give expression embeddings of micro-expression images which are regarded as false labels. We introduce a discriminator D to identify these two embeddings. Our MicroNet aims to generate micro-expression embeddings that the discriminator can not distinguish from macro-expression embeddings with same expression labels; while the discriminator aims to distinguish between these two kinds of embeddings. Through adversarial learning, MicroNet can be fine tuned to model the shared features of micro-expression and macro-expression images. The objective of our adversarial learning is thus as:

$$\min_{MicroNet} \max_{D} \mathbb{E}_{f^e_{pos} \sim P(f^e_{pos})} \log D(f^e_{pos}) + \mathbb{E}_{f^e_{anc} \sim P(f^e_{anc})} \log\{1 - D(f^e_{anc})\} \quad (7)$$

As shown in Goodfellow *et al.*'s work [6], the above equation can not been optimized directly, thus the loss of discriminator is defined as:

$$\mathcal{L}_{\mathcal{D}} = -\log D(f^e_{pos}) - \log\{1 - D(f^e_{anc})\} \quad (8)$$

It is better to minimize $-\log D(f^e_{anc})$ instead of minimizing $\log\{1 - D(f^e_{anc})\}$ in order to avoid flat gradients, thus the adversarial loss of MicroNet is defined as:

$$\mathcal{L}_{adv} = -\log D(f^e_{anc}) \quad (9)$$

3.2.2 Guidance in Label Space. The classification loss is used to control recognition accuracy:

$$\mathcal{L}_{cls} = -\sum_{l=1}^{L} 1_{[y=l]} \log(D_e(f^e_{anc})) \quad (10)$$

where y is the expression label of x_{anc}, and $D_e(f^e_{anc})$ is predicted expression class distribution.

Up to now, the guidance of macro-expression merely happens in feature space. There is a lack of guidance in label space. During training, we jointly train MicroNet and MacroNet by assuming these two networks produce similar outputs. For that purpose, we think about adding a regularization term in the loss function to penalize the differences of two networks. Motivated by Wang *et al.* 's work [36], we introduce a regularization method called loss inequality regularization (LIR). This method is based on the assumption that secondary feature is more informative than the primary feature.

Database		SMIC[16]	CASME2[37]	SAMM[2]	3DB-combined
Subjects		16	24	28	68
Emotions	Negative	70	88	92	250
	Positive	51	32	26	109
	Surprise	43	25	15	83
	Total	164	145	133	442

Table 1: 3-class sample distribution of all databases for CDE task.

In our case, macro-expression images are more informative than micro-expression images. The basic idea is to penalize the violation of this constraint, the LIR loss is defined as:

$$\mathcal{L}_{LIR} = max\{\mathcal{L}_{cls} - \mathcal{L}_{cls'}, 0\} \quad (11)$$

where $\mathcal{L}_{cls'}$ is the cross entropy between the result of classifier onto positive feature f^e_{pos} and the true label of positive images y.

Thus the overall loss of MTMNet is:

$$\mathcal{L}_{MTM} = \mathcal{L}_{cls} + \lambda_{2,1}\mathcal{L}_{tri} + \lambda_{2,2}\mathcal{L}_{adv} + \lambda_{2,3}\mathcal{L}_{LIR} \quad (12)$$

where $\lambda_{2,1}$, $\lambda_{2,2}$ and $\lambda_{2,3}$ are the hyperparameters controlling loss coefficients.

4 EXPERIMENTS

4.1 Experiments Condition

As the micro-expression community always uses CASME2 [37], SMIC [16] and SAMM [2] databases as evaluation standards for recognition tasks [24, 31], we adopt this custom in our paper.

The CASME2 dataset has 249 micro-expressions from 26 subjects, with the average age of 22.03 years old at 200 fps. The resolution of the samples is 640 × 480 pixels and the resolution of face area is around 280 × 340 pixels. The CASME2 dataset includes five micro-expression classes, i.e., happiness, surprise, disgust, repression and others.

The SMIC dataset contains 164 micro-expression samples from 16 participants. The frame rate is 100 fps. The resolution is 640 × 640 pixels, and the resolution of the face area is around 190 × 230 pixels. There are three micro-expression types in the SMIC dataset, including negative, positive and surprise.

The SAMM dataset contains 159 micro-expression clips from 32 participants at 200 fps. These participants are from 13 races and the average age is 33.24 years old. The resolution of the samples is 2040 × 1088 pixels and the resolution of face area is around 400 × 400 pixels. Seven micro-expression types are included in the SAMM dataset. They are happiness, surprise, disgust, repression, angry, fear and contempt.

We mainly conduct two experiments on these databases. First, we test our proposed framework on the CASME2, SMIC and SAMM databases separately. Second, we test our framework on Composite Database Evaluation (CDE) task [24, 31], i.e., samples from all databases are combined into a single composite database based on the reduced emotion classes. Specifically, the samples of happiness are given positive labels and the labels of surprise samples are unchanged. The samples of disgust, repression, anger, contempt, fear and sadness are categorized into negative. The distribution of samples and subjects for CDE task are given in Table 1. We also take ablation study on our proposed EIDNet loss and MTMNet loss

Method		CASME2(CK+)		SMIC(CK+)		SAMM(CK+)		CASME2(MMI)		SMIC(MMI)		SAMM(MMI)		CASME2(Oulu)		SMIC(Oulu)		SAMM(Oulu)	
EIDNet Loss	MTMNet Loss	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
-	$\mathcal{L}_{cls}$	0.635	0.567	0.604	0.591	0.632	0.624	0.635	0.567	0.604	0.591	0.632	0.624	0.635	0.567	0.604	0.591	0.632	0.624
-	$\mathcal{L}_{cls} + \mathcal{L}_{tri}$	0.670	0.600	0.659	0.646	0.661	0.673	0.677	0.611	0.675	0.675	0.665	0.670	0.691	0.632	0.655	0.670	0.673	0.675
-	$\mathcal{L}_{cls} + \mathcal{L}_{adv}$	0.675	0.607	0.646	0.635	0.670	0.678	0.671	0.615	0.682	0.682	0.685	0.655	0.682	0.650	0.675	0.645	0.682	0.661
-	$\mathcal{L}_{cls} + \mathcal{L}_{tri} + \mathcal{L}_{adv}$	0.682	0.612	0.671	0.661	0.672	0.681	0.685	0.621	0.685	0.678	0.685	0.680	0.691	0.655	0.680	0.675	0.685	0.678
-	$\mathcal{L}_{cls} + \mathcal{L}_{LIR}$	0.658	0.590	0.634	0.622	0.661	0.664	0.661	0.594	0.654	0.638	0.654	0.635	0.672	0.641	0.634	0.625	0.664	0.648
-	$\mathcal{L}_{MTM}$	0.691	0.633	0.683	0.676	0.685	0.691	0.691	0.633	0.701	0.695	0.701	0.695	0.703	0.668	0.701	0.695	0.698	0.695
$\mathcal{L}_{EID}$	$\mathcal{L}_{cls} + \mathcal{L}_{tri}$	0.711	0.654	0.732	0.717	0.715	0.718	0.734	0.661	0.722	0.730	0.715	0.718	0.664	0.550	0.705	0.718	0.718	0.718
$\mathcal{L}_{EID}$	$\mathcal{L}_{cls} + \mathcal{L}_{adv}$	0.723	0.666	0.720	0.707	0.725	0.711	0.715	0.683	0.732	0.715	0.723	0.711	0.684	0.550	0.722	0.711	0.723	0.711
$\mathcal{L}_{EID}$	$\mathcal{L}_{cls} + \mathcal{L}_{tri} + \mathcal{L}_{adv}$	0.731	0.683	0.738	0.725	0.727	0.720	0.736	0.686	0.732	0.732	0.725	0.718	0.743	0.689	0.725	0.718	0.721	0.722
$\mathcal{L}_{EID}$	$\mathcal{L}_{cls} + \mathcal{L}_{LIR}$	0.699	0.638	0.707	0.697	0.709	0.697	0.705	0.656	0.711	0.701	0.703	0.702	0.712	0.675	0.703	0.697	0.701	0.701
$\mathcal{L}_{EID}$	$\mathcal{L}_{MTM}$	**0.743**	**0.693**	**0.750**	**0.742**	**0.741**	**0.736**	**0.739**	**0.688**	**0.768**	**0.744**	**0.738**	**0.732**	**0.756**	**0.701**	**0.738**	**0.726**	**0.734**	**0.734**

Table 2: Accuracy and F1 Score results on the CASME2, SMIC and SAMM databases separately. CK+ denotes using the CK+ database as macro-expression database. MMI denotes using the MMI database as macro-expression database. Oulu denotes using the Oulu-CASIA database as macro-expression database.

in all experiments. Leave-one-subject-out (LOSO) cross-validation is used in all experiments. In order to compare with related works, for the first experiment, accuracy and F1 score are used for evaluations. For the second experiment, unweighted F1 score (UF1) and unweighted accuracy (UAR) are adopted.

Since the proposed method requires both neutral facial images and expression facial images of same subject, wild-collected databases, i.e. AffectNet [25] and RAF-DB [14] databases are not suitable. Three popular lab-collected databases, i.e., CK+[23], MMI[26] and Oulu-CASIA[38] database are adopted. The CK+ database is composed of 327 image sequences of seven emotion labels: anger, contempt, disgust, fear, happiness, sadness, and surprise. The Oulu-CASIA database includes 480 image sequences of six emotion labels: anger, disgust, fear, happiness, sadness, or surprise. The MMI database consists of 205 image sequences with frontal faces of six emotion labels: anger, disgust, fear, happiness, sadness and surprise. Since we need macro-expression images that have the same label with micro-expression images, only related parts of macro-expression database are used in each experiment corresponding to each micro-expression database.

We do not use all frames on the micro-expression databases, since many frames contain little or no additional information to neutral frames according to the brevity of micro expressions. In CASME2, SMIC and SAMM databases, only five frames centered at the apex frame are chosen for experiments. In our experiments, all facial images are resized to 224×224 pixels.

We choose ResNet18 as backbone of the encoder since more complicated structures like ResNet34 and ResNet101 only raise little performance and lighter structures such as AlexNet would result in a noticeable fall in the performance. Two branches are linked after the backbone and extract expression and identity embeddings separately. For decoder, it uses up-sampling to double feature map size and also implements convolutional layers with ReLU and Batch normalization. The structure of discriminator is that several convolutional layers ending with a linear layer outputs a scalar value.

When training the MicroNet and MacroNet, we set $\lambda_{1,1} = \lambda_{1,2} = \lambda_{1,3} = 0.1$. The learning rates of the encoder, classifier and decoder are set to 10^{-4}, 10^{-4} and 10^{-5} separately. When training the MTMNet, $\lambda_{2,1} = \lambda_{2,2} = \lambda_{2,3} = 10^{-3}$ and the learning rates of the encoder, discriminator and the classifier are set to 10^{-5}, 10^{-5} and 10^{-5}. Every fold of LOSO procedure is trained with total 20 epochs.

Method		CK+		MMI		Oulu-CASIA	
EIDNet Loss	MTMNet Loss	UF1	UAR	UF1	UAR	UF1	UAR
-	$\mathcal{L}_{cls}$	0.685	0.683	0.685	0.683	0.685	0.683
-	$\mathcal{L}_{cls} + \mathcal{L}_{tri}$	0.749	0.735	0.762	0.744	0.756	0.745
-	$\mathcal{L}_{cls} + \mathcal{L}_{adv}$	0.734	0.753	0.742	0.761	0.734	0.760
-	$\mathcal{L}_{cls} + \mathcal{L}_{tri} + \mathcal{L}_{adv}$	0.756	0.759	0.766	0.764	0.756	0.766
-	$\mathcal{L}_{cls} + \mathcal{L}_{LIR}$	0.718	0.728	0.728	0.731	0.730	0.731
-	$\mathcal{L}_{MTM}$	0.786	0.787	0.795	0.793	0.776	0.780
$\mathcal{L}_{EID}$	$\mathcal{L}_{cls} + \mathcal{L}_{tri}$	0.838	0.821	0.841	0.821	0.830	0.815
$\mathcal{L}_{EID}$	$\mathcal{L}_{cls} + \mathcal{L}_{adv}$	0.817	0.829	0.828	0.840	0.818	0.829
$\mathcal{L}_{EID}$	$\mathcal{L}_{cls} + \mathcal{L}_{tri} + \mathcal{L}_{adv}$	0.845	0.842	0.845	0.842	0.838	0.829
$\mathcal{L}_{EID}$	$\mathcal{L}_{cls} + \mathcal{L}_{LIR}$	0.805	0.798	0.815	0.822	0.805	0.810
$\mathcal{L}_{EID}$	$\mathcal{L}_{MTM}$	**0.870**	**0.856**	**0.862**	**0.858**	**0.848**	**0.842**

Table 3: UF1 and UAR results of CDE task with different macro-expression databases.

4.2 Experimental Results and Analysis

4.2.1 The Effect of EIDNet Loss. In order to evaluate the effect of our proposed EIDNet, we compare the method using EIDNet as feature encoder with the method does not using EIDNet. As shown in Table 2, when we do not use $\mathcal{L}_{EID}$ to pretrain EIDNet, our proposed MTMNet just gains 5.6%/6.6%, 7.9%/8.5% and 5.3%/6.7% increases in accuracy/F1 score than the baseline model that only using classification loss $\mathcal{L}_{cls}$ on the CASME2, SMIC and SAMM database, by using CK+ as macro-expression database. When adopting $\mathcal{L}_{EID}$ to pretrain feature encoder, our method gains 10.8%/12.6%, 14.6%/15.1% and 10.9%/11.2% increases than the baseline on the CASME2, SMIC and SAMM database. It means the introduction of EIDNet can improve the performance of MTMNet by obtaining expression-related embeddings. No matter which macro-expression database we use for assisting, the MTMNet all gains incredible increases with EIDNet as feature encoder in the CDE task. As shown in Table 3, we find the MTMNet with EIDNet outperforms the MTMNet without EIDNet by 8.4%/6.9%, 6.7%/6.5% and 7.2%/6.2% of UF1/UAR, through the guidance of CK+, MMI and Oulu-CASIA database respectively. Since subjects in macro-expression databases and micro-expression databases are different, training MTMNet will suffer from identity-related features distortion. However, our proposed

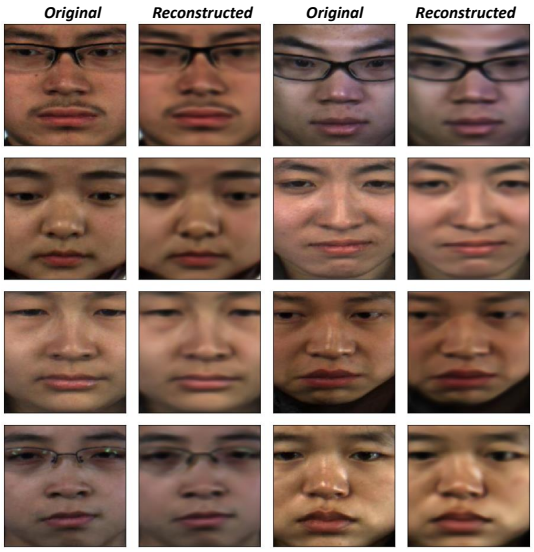

Figure 3: Comparison between the original micro-expression images and the reconstructed images on the CASME2 database. The first and third columns are original images, the second and fourth columns are corresponding reconstructed images.

framework adopting EIDNet as feature extractor can avoid this disadvantage and take full use of vast amounts of macro-expression images.

4.2.2 Visualization of Reconstructed Images. As elaborated in Sec 3.1, we introduce a reconstruction loss in the EIDNET to get original expression facial images from concatenating identity-related and expression-related features. Thus, we visualize the reconstructed facial images to see what the EIDNET learns. Taking CASME2 database as example, we visualize the original images and the reconstructed images in Figure 3. From Figure 3, we can see that there is nearly no difference between the original facial images and the reconstructed facial images, indicating that the EIDNet separates the identity-related and expression-related features effectively.

4.2.3 The Effect of MTMNet Loss. We conduct ablation experiments to verify the influence of three different loss functions in the MTMNet, i.e., triplet loss, adversarial loss and LIR loss on the final recognition performance. As shown in Table 2 and 3, We can draw the following observations:

Firstly, the guidance from the feature and label spaces all lead to a great improvement of micro-expression recognition accuracy and f1 score. For example, the UF1/UAR of $\mathcal{L}_{cls} + \mathcal{L}_{tri}$, $\mathcal{L}_{cls} + \mathcal{L}_{adv}$ and $\mathcal{L}_{cls} + \mathcal{L}_{LIR}$ are 15.3%/13.8%, 13.2%/14.6% and 12.0%/11.5% higher than the baseline that only using $\mathcal{L}_{cls}$ on the CDE task, by taking CK+ database as guidance. Even if we don't use EIDNet as feature encoder, MTMNet can also improve the performance of micro-expression recognition through the guidance from the feature and label spaces.

Method	CASME2		SMIC		SAMM	
	Acc	F1	Acc	F1	Acc	F1
LBP-TOP [13]	0.490	0.510	0.580	0.600	0.590	0.364
LBP-SIP [35]	0.465	0.448	0.445	0.449	0.415	0.406
STLBP-IP [8]	0.595	0.570	0.579	0.580	0.568	0.527
HIGO [15]	0.672	-	0.682	-	-	-
FHOFO [7]	0.566	0.524	0.518	0.524	-	-
Bi-WOOF [20]	0.588	0.610	0.622	0.620	0.583	0.397
STCLQP [9]	0.640	0.638	0.583	0.583	0.638	0.611
Only-Apex [17]	0.633	-	-	-	-	-
OFF-Apex [5]	-	-	0.676	0.670	0.681	0.542
CNN+LSTM [12]	0.609	-	-	-	-	-
Boost [28]	0.709	-	0.689	-	-	-
DSSN [11]	0.708	**0.730**	0.634	0.646	0.574	0.464
Dynamic [32]	0.726	0.670	0.761	0.710	-	-
ours	**0.756**	0.701	**0.768**	**0.744**	**0.741**	**0.736**

Table 4: Comparison with state-of-the-art methods on the CASME2, SMIC and SAMM databases separately.

Method	Full		SMIC		CASME2		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [39]	0.588	0.578	0.200	0.528	0.702	0.742	0.395	0.410
Bi-WOOF [20]	0.629	0.622	0.572	0.582	0.780	0.802	0.521	0.513
OFF-Apex [5]	0.719	0.709	0.681	0.669	0.876	0.868	0.540	0.539
Capsule [34]	0.652	0.650	0.582	0.587	0.706	0.701	0.620	0.598
Shallow [19]	0.735	0.760	0.680	0.701	0.838	0.868	0.658	0.681
Dual [40]	0.732	0.727	0.664	0.672	0.862	0.856	0.586	0.566
Neural [21]	0.788	0.782	0.746	0.753	0.829	0.820	0.775	0.715
ours	**0.864**	**0.857**	**0.864**	**0.861**	**0.870**	**0.872**	**0.825**	**0.819**

Table 5: Comparison with state-of-the-art methods of CDE task.

Secondly, the triplet loss and adversarial loss can lead to different promotions for micro-expression recognition. The triplet loss greatly improves the recognition F1 score, while the adversarial loss gains recognition accuracy increases. In the CDE task, $\mathcal{L}_{cls} + \mathcal{L}_{tri}$ is 15.3%, 15.6% and 14.5% higher than the baseline in UF1, but only obtains 13.8%, 13.8% and 13.2% increases in UAR, with CK+, MMI and Oulu-CASIA database as guidacne respectively. On the contrary, $\mathcal{L}_{cls} + \mathcal{L}_{adv}$ is 14.6%, 15.7% and 14.6% higher than the baseline in UAR, but only obtains 13.2%, 14.3% and 13.3% increases in UF1, with CK+, MMI and Oulu-CASIA database as guidacne respectively. When we combine the triplet loss and adversarial loss from feature space, the MTMNet greatly improves the accuracy and f1 score at the same time. To be specific, $\mathcal{L}_{cls} + \mathcal{L}_{tri} + \mathcal{L}_{adv}$ outperforms the baseline by 16.0%, 16.0% and 15.3% in UF1, 15.9%, 15.9% and 14.6% in UAR through the guidance of CK+, MMI and Oulu-CASIA database.

Finally, our method combines the strengths of the three introduced loss functions and achieves the best performance. Specifically, the UF1/UAR of our method is 18.5%/17.3%, 17.7%/17.5% and 16.3%/15.9% higher than the baseline on the CDE task, with CK+, MMI and Oulu-CASIA database as guidacne respectively, which is much better than using a single loss functions. This indicates that the different guidance will not cause the inter-domian discrepancy. Guidance in both feature and label spaces can help MTMNet learn more robust feature representations and make better predictions

4.2.4 Analysis of Macro-expression Databases. In order to analyze what kind of macro-expression database is more conducive to assisting the training of micro-expression classifier, we choose CK+, MMI and Oulu-CASIA as guidance respectively. As shown in Table 2 and Table 3, we can get the following observations:

Firstly, the guidance from Oulu-CASIA database achieves the best results on the CASME2 database, but get the worst results on the SAMM and SMIC database, compared with the other two macro-expression databases. When we evaluate the subject distribution of all databases, we find that this may caused by dataset bias. The CASME2 and Oulu-CASIA database both contain predominantly Chinese subjects, while other databases have balanced distribution of nationalities. The images of CASME2 all are collected from Chinese, and the Oulu-CASIA database has a total of 80 subjects, including 30 Chinese. This indicates micro-expression and macro-expression may have greater similarity within the same nationality than different nationalities, greater similarity can lead to better enhancement for micro-expression classifier.

Secondly, since composite micro-expression database contains a diverse range of age and ethnicity, we require diverse macro-expression images as guidance. Because the CK+ and MMI database contain diverse samples, choosing these two databases as guidance can achieve better performance on the overall CDE task. Specifically, using CK+ database as macro-expression database can get higher UF1 score, while using MMI database as macro-expression database can get higher UAR score. This guides us to using macro-expression database of diversity distribution as guidance to improve the generalization of our micro-expression classifier.

4.3 Comparison with Related Works

We compare our framework with other related works. These methods are: 1) LBP-TOP [13], LBP-SIP [35], STLBP-IP [8], STCLQP [9], which are LBP based methods 2) HIGO [15], FHOFO [7], Bi-WOOF [20], which are optimal flow based methods, 3) Only-Apex [17], OFF-Apex [5], CNN+LSTM [12], Boost [28], DSSN [11], Shallow [19], Dual [40] Capsule [34], which are deep feature methods, 4) Dynamic [32], which is action unit assisted method. and 5) Neural [21], which is macro-expression assisted method.

4.3.1 Experiments on the CASME2, SAMM and SMIC Database Separately. From Table 4, we can see that our framework exceeds most handcraft feature methods, i.e., LBP and optimal flow based methods in almost every evaluation indicators. Our framework achieves nearly 16.8%, 14.6% and 15.8% increases in accuracy, 9.1%, 12.4% and 33.9% increases in f1 score compared to the best results of handcraft feature methods, i.e., Bi-WOOF [20] on the CASME2, SAMM and SMIC database. Our method also outperforms the action unit assisted method, i.e., Dynamic [32] by 3.0%/3.1% and 0.7%/3.4% on the CASME2 and SMIC database of accuracy/F1. Since Dynamic enforces the features of teacher network ans student network similar by L2-loss, student network would lose its own domain information on micro-expression recognition. While, in this paper we introduce adversarial learning and triplet loss to make micro-expression and macro-expression images produce similar feature distributions.

4.3.2 Experiments of CDE Task. Our method gains higher results compared to state-of-the-art deep feature methods, i.e., Shallow [19], Dual [40] and Capsule [34] on CDE task. Because the CDE task has greatly increased micro-expression training data, these deep feature methods can't handle the difference between databases very well. As a result, these methods perform well on the CASME2 database, but get poor results on the SMIC and SAMM database. However, our proposed framework adopts EIDNet as feature extractor to avoid identity-related features distortion, and take full use of composite database. Our method only exceeds Shallow and Dual by 3.2%/0.4% and 0.8%/1.6% in CAMSE2 database, but gain incredible increases than them by 18.4%/16.0% and 20.0%/18.9% on the SMIC database, 16.7%/13.8% and 23.9%/25.3% on the SAMM database of UF1/UAR.

Compared with the macro-expression assisted method, i.e., Neural [21], our proposed framework exceeds the results of it in full or separated single databases, with 7.6%/7.5% on full database, 11.8%/10.8% on the SMIC database, 4.10%/5.20% on the CAMSE2 database and 5.0%/10.4% on the SAMM database of UF1/UAR. Neural used Expression Magnification and Reduction (EMR) to reduce the gap between micro and macro expression. This preprocessing causes micro and macro expression visually similar, but this can't guarantee the similarity of micro and macro expression features, then directly training on a fusion of micro and macro expression database can't generate appropriate expression feature representation. Our method use adversarial learning and triplet loss to fine-tuning MicroNet, which can model the shared features of micro-expression and macro-expression samples effectively.

5 CONCLUSION

Our paper presents a micro-expression recognition method by leveraging macro-expression databases as guidance. Expression-Identity Disentangle Network is also proposed to extract expression embeddings from expression image without identity-related information. Extensive experiments on the three public spontaneous micro-expression databases, i.e., SMIC, CASME2 and SAMM demonstrated that our framework outperformed the state-of-the-art micro-expression recognition methods based on either handcraft or deep features. And for future research routines, our proposed method also cast a light into a research direction of combining micro and macro expression recognition problems.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China 61727809, 91741812.

REFERENCES

[1] Xianye Ben, Xitong Jia, Rui Yan, Xin Zhang, and Weixiao Meng. 2018. Learning effective binary descriptors for micro-expression recognition transferred by macro-information. *Pattern Recognition Letters* 107 (2018), 50–58.
[2] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2016. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing* 9, 1 (2016), 116–129.
[3] Paul Ekman. 2009. Lie catching and microexpressions. *The philosophy of deception* (2009), 118–133.
[4] Paul Ekman. 2009. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition).* WW Norton & Company.
[5] YS Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. 2019. OFF-ApexNet on micro-expression recognition system. *Signal Processing: Image Communication* 74 (2019), 129–139.
[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[7] SL Happy and Aurobinda Routray. 2017. Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Transactions on Affective Computing* (2017).

[8] Xiaohua Huang, Su-Jing Wang, Guoying Zhao, and Matti Piteikainen. 2015. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In *Proceedings of the IEEE international conference on computer vision workshops*. 1–9.

[9] 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.

[10] Xitong Jia, Xianye Ben, Hui Yuan, Kidiyo Kpalma, and Weixiao Meng. 2018. Macro-to-micro transformation model for micro-expression recognition. *Journal of Computational Science* 25 (2018), 289–297.

[11] Huai-Qian Khor, John See, Sze-Teng Liong, Raphael CW Phan, and Weiyao Lin. 2019. Dual-stream shallow networks for facial micro-expression recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 36–40.

[12] 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 24th ACM international conference on Multimedia*. ACM, 382–386.

[13] Anh Cat Le Ngo, John See, and Raphael C-W Phan. 2016. Sparsity in dynamics of spontaneous subtle emotions: analysis and application. *IEEE Transactions on Affective Computing* 8, 3 (2016), 396–411.

[14] Shan Li, Weihong Deng, and JunPing Du. 2017. 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*. 2852–2861.

[15] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. 2017. 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.

[16] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. 2013. 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, 1–6.

[17] Yante Li, Xiaohua Huang, and Guoying Zhao. 2018. Can micro-expression be recognized based on single apex frame?. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 3094–3098.

[18] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2019. Self-Supervised Representation Learning From Videos for Facial Action Unit Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10924–10933.

[19] Sze-Teng Liong, YS Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. 2019. 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, 1–5.

[20] 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 Processing: Image Communication* 62 (2018), 82–92.

[21] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. 2019. A neural micro-expression recognizer. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–4.

[22] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. 2015. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing* 7, 4 (2015), 299–310.

[23] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. 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, 94–101.

[24] Walied Merghani, Adrian Davison, and Moi Yap. 2018. Facial Micro-expressions Grand Challenge 2018: evaluating spatio-temporal features for classification of objective classes. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 662–666.

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

[26] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*. IEEE, 5–pp.

[27] 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 *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 657–661.

[28] Wei Peng, Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. 2019. A boost in revealing subtle facial expressions: A consolidated eulerian framework. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–5.

[29] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. 2012. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*. Springer, 808–822.

[30] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in neural information processing systems*. 3856–3866.

[31] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. 2019. Megc 2019–the second facial micro-expressions grand challenge. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–5.

[32] Bo Sun, Siming Cao, Dongliang Li, Jun He, and Lejun Yu. 2020. Dynamic Micro-Expression Recognition Using Knowledge Distillation. *IEEE Transactions on Affective Computing* (2020).

[33] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1415–1424.

[34] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. 2019. Capsulenet for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–7.

[35] 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 *Asian conference on computer vision*. Springer, 525–537.

[36] Ziheng Wang and Qiang Ji. 2015. Classifier learning with hidden information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4969–4977.

[37] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one* 9, 1 (2014), e86041.

[38] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (2011), 607–619.

[39] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6 (2007), 915–928.

[40] Ling Zhou, Qirong Mao, and Luoyang Xue. 2019. Dual-inception network for cross-database micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–5.