Learning from Macro-expression: an Adversarial Micro-expression Recognition Framework

Anonymous CVPR submission

Paper ID 9022

Abstract

In this paper, we aim to facilitate a deep neural network to do micro-expression recognition (MEF) task with high accuracy. Inspired by the fact that micro-expressions are indeed subtle and very brief facial expressions, we adopt an adversarial learning framework that leverages macro-expression databases as privileged information to assist. Due to the fact that the subjects in micro and macro expression databases are different, we introduce the Expression-Identity Disentangle Network (EIDNet) as the feature extractor to disentangle expression-related features apart from identity-related features. Comprehensive experiments on the three public spontaneous micro-expression databases SMIC, CASME2 and SAMM show that our method is state-of-the-art.

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 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 as very brief, subtle, and involuntary facial expressions which normally occur when a person either deliberately or unconsciously conceals his or her genuine emotions [4, 3]. It always takes human beings lots of time to perceive and recognize them. Thus developing micro-expression recognition systems becomes necessary. Figure 1 shows a comparison between micro and maco expressions.

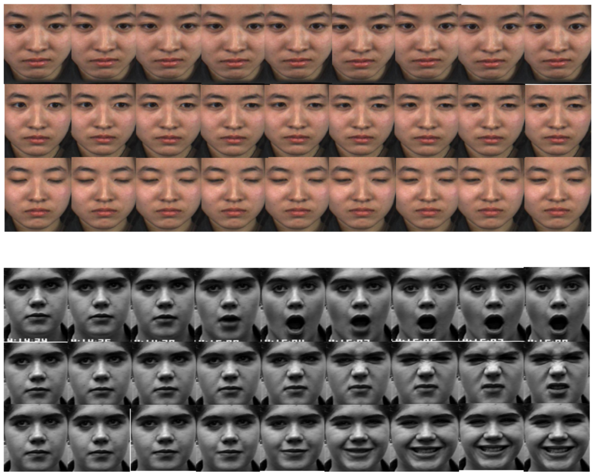

Figure 1. Here we give examples of micro-expression and macro-expression, where the above three rows are from CASME2 database and the below three rows are from CK+ database, both with surprising, disgusting and happiness expression labels sequentially. Nine frames are randomly chosen from each micro-expression or macro-expression video sequentially for each row We can find it's easy to detect expression changes for macro-expressions, while expression differences in micro-expression sequence is very subtle and take time to carefully identity.

There are lots of efforts devoted to micro-expression recognition, which fall into two main kinds: handcraft feature methods [13, 20, 18, 9, 8, 32, 7] and deep feature methods [11, 5, 15, 28, 36, 31, 19, 17]. However, though easily implementing and embracing good geometric or spatiotemporal interpretations, handcraft features are actually too 'shallow' for extracting intrinsic features of the subtle and brief micro-expressions. As for deep networks, though powerful, they are limited by the scarcity of micro-expression databases. Only enough data points can we use to implement efficient deep network with good generalization ability.

In order to address problems mentioned above, we propose an adversarial micro-expression recognition framework that leverages macro-expression, i.e., normal facial expression, databases as privilege information to assist. Since subjects in macro-expression and micro-expression databases are different, an Expression-Identity Disentangle

Network (EIDNet) is introduced as feature extractor.

In summary, our contributions are three-folds: 1) an adversarial learning framework is proposed for micro-expression recognition by leveraging macro-expression databases to assist. To the best of our knowledge, it's the first time adversarial learning methods are used in micro-expression recognition with macro-expression databases. 2) We use an Expression-Subject Disentanglement Network (EIDNet) to disentangle expression-related features apart from subject-related features of either micro-expression or macro-expression images to let assistant across micro and macro expression databases become possible. 3) Extensive experiments on three micro-expression databases show the superiority of our framework compared to other state-of-the-art micro-expression recognition protocols, which illuminates a new direction into combination of micro-expression and macro-expression recognition.

2. Related Work

Micro-Expression Recognition. Micro-expression recognition methods can be categorized into two main kinds as handcrafted feature methods and deep feature methods.

Histogram of gradient (HOG), Optimal flow and Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) based methods are the most prevailing handcraft feature methods. [13] adopted a variant of HOG in their work. For optical flow based methods, [20] proposed the Main Directional Mean Optical-flow (MDMO) method to describe micro-expressions and showed its superiority against LBP-TOP and HOOF features. [18] adopted a Bi-Weighted Oriented Optical Flow (Bi-WOOF) based feature extractor, while [7] proposed a fuzzy HOFO methods. LBP-TOP methods [35] and its variants gained the most popularity. Huang *et al.* used LBP-TOP methods in their early works [9, 8]. Wang *et al.* adopted a pruned LBP descriptor using six neighbors around every point [32]. Our paper mainly focus on framework based on deep network, thus we left readers the complete and further introduction of all handcrafted feature methods to [24, 25]. Many researchers turn to the deep networks as the micro-expression databases gradually developed. In [11], their two-phases method 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.* proposed a CNN framework with optimal flow between the apex frame and onset frame as input features [5]. [15] used only apex frame passing through a deep neutral network to get features. Nguyen *et al.* adopted the newly proposed framework CapsuleNet [28] to recognize micro-expressions [31], while Ling *et al.* proposed a dual inception networks taking horizontal and vertical optimal flow features as inputs [36]. [17] aimed to put up with a compressed deep architecture with optimal flow features as inputs as well.

Recent years, there have already been some works embracing the ideas of using macro-expression databases as privileged information for micro-expression recognition. [10] and [26] both adopted transfer learning protocols with different details, where [26] trained a micro-expression recognition network pretrained on macro-expression database and [10] combined features extracted from macro-expression images by LBP descriptor with features extracted from micro-expression images by LBP-TOP descriptor during training process of micro-expression recognition network. In [1], coupled metric learning algorithm was employed to model the shared features between micro-expression and macro-expression samples. Liu *et al.* [19] proposed a micro-expression recognition network by training on a fusion of micro and macro expression databases. However, none of them adopt adversarial learning protocol between micro-expression and macro-expression databases as in our method.

Feature Disentanglement. Feature disentanglement technique aims to disentangle different kinds of features from original inputs for specific uses, which can help raising performance and implementing complicated model since more domain related features are provided. We focus on related disentangle works in expression recognition areas. There are two main directions: 1) disentangle facial expression apart from pose or head motions. 2) disentangle facial expression apart from identity information. In [16], Li *et al.* proposed a self-supervised disentangle auto-encoder for distinguishing AU-related features from motion-related features of facial expression images. Tran *et al.* learned a disentangled representation learning-generative adversarial network (DR-GAN) for learning facial expression apart from pose variances [30]. In [27], Salah *et al.* used features representation produced by the multi-scale contractive convolutional network (CCNET) to train a Contractive Discriminative Analysis (CDA) feature extractor for learning a representation separating out the emotion-related factors from the others (which mostly capture the subject identity, and what is left of pose after the CCNET). In our paper, We use feature disentanglement techniques to get expression embeddings apart from identity embeddings in both micro-expression and macro-expression databases. Only in this way can we efficiently leverage adversarial learning procedure across databases.

3. Problem Formulation

Our goal is to train a deep network for micro-expression recognition task with macro-expression databases as privileged information by adversarial learning methods. A recognizer with high accuracy is our final goal, $\mathcal{F} : \mathbb{R}^{H \times W \times 3} \longrightarrow \mathbf{0}, \mathbf{1}, ..., \mathbf{K-1}$. The training process of $\mathcal{F}$ splits into two phases. Firstly, $\mathcal{D}_{train} = \{I_N, I_E, y\}$, where I_N, I_E are neutral and expression images of a same

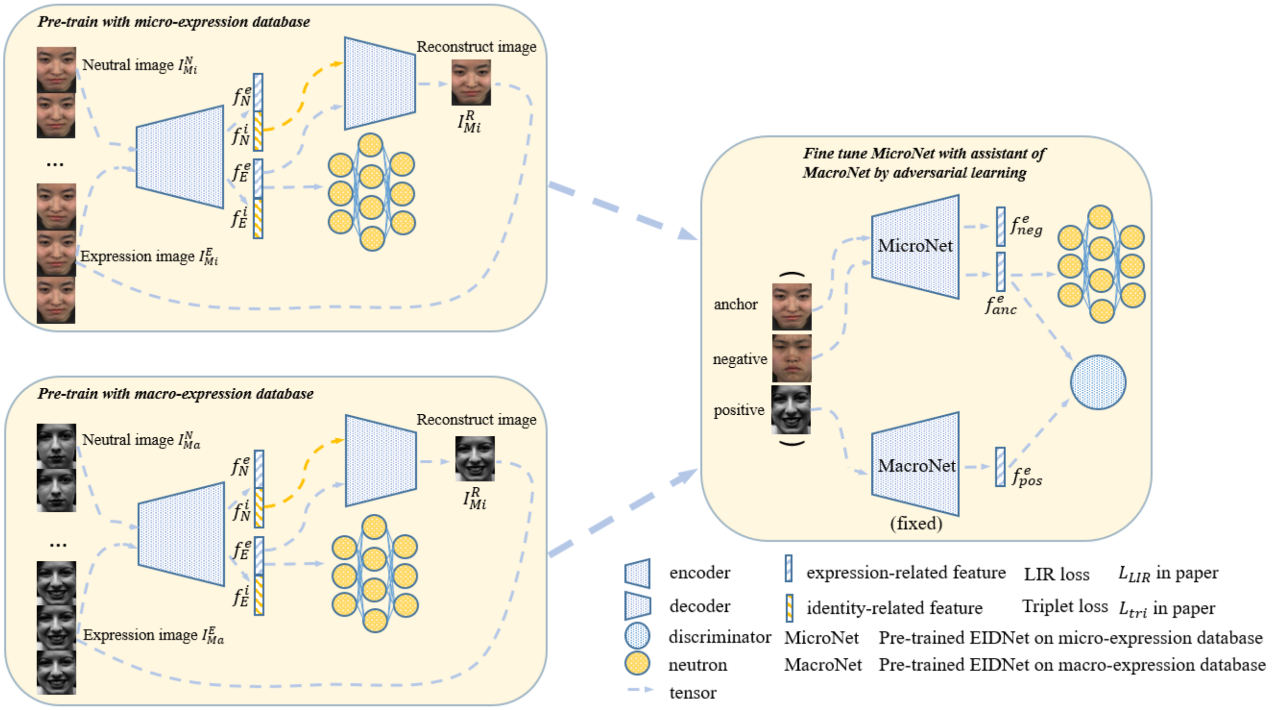
Pre-train with micro-expression database
Neutral image I^N_{Mi}
...
Expression image I^E_{Mi}
f^e_N
f^i_N
f^e_E
f^i_E
Reconstruct image
I^R_{Mi}
Pre-train with macro-expression database
Neutral image I^N_{Ma}
...
Expression image I^E_{Ma}
f^e_N
f^i_N
f^e_E
f^i_E
Reconstruct image
I^R_{Mi}
Fine tune MicroNet with assistant of MacroNet by adversarial learning
anchor
negative
positive
MicroNet
MacroNet
(fixed)
f^e_{neg}
f^e_{anc}
f^e_{pos}
encoder
decoder
discriminator
neutron
tensor
expression-related feature
identity-related feature
LIR loss L_{LIR} in paper
Triplet loss L_{tri} in paper
MicroNet Pre-trained EIDNet on micro-expression database
MacroNet Pre-trained EIDNet on macro-expression database

Figure 2. The framework of our recognition system. First we pre-train two EIDNets with micro-expression and macro-expression databases separately, named MicroNet and MacroNet. Feature disentangle techniques are adopted for extracting expression embeddings apart from identity embeddings. Secondly, an adversarial learning module was added onto the feature level between these two pre-trained nets to fine tune MicroNet.

video from micro-expression database and y is the corresponding label. A same network trained with macro-expression database is also offered, named as $\mathcal{F}_{privileged}$. Secondly, $\mathcal{D}_{train} = \{I_{anc}, I_{pos}, I_{neg}, y\}$, where I_{anc}, I_{neg} are micro-expression images with different labels, while I_{pos} is macro-expression image labeled the same as I_{anc} and y is the label. I_{anc}, I_{neg} will go through $\mathcal{F}$ while I_{pos} goes through $\mathcal{F}_{privileged}$. Structures will be added to further fine tune $\mathcal{F}$.

4. Our Method

As mentioned in section 3, our framework is with two consecutive training phases. In the first phase, two expression-identity disentangle networks (EIDNets) are pre-trained with micro-expression and macro expression databases separately. In the second phase, adversarial learning module is added upon feature level between two pre-trained networks for further fine tuning. Figure 2 gives a full illustration. Below we will give clarifications of framework structures according to different phases. The details of training will be given in section 5.

4.1. Structure of EIDNets

The inputs of each EIDNet are pairs of images derived from the same video, consisting of a neutral facial image I_N and an expression facial image I_E. EIDNet mainly consists of three parts, named feature disentangling, expression reconstruction and classification. In feature disentangling, EIDNet learns the features by respectively extracting the identity-related and expression-related features between two images throught the two-branches encoder. In expression reconstruction, EIDNet integrates the identity feature of neutral facial image and the expression feature of expression facial image and uses it to reconstruct the expression facial image, ensuring that these two features are sufficient to represent the information of the inputs. Due to the differences between micro-expression and facial expression recognition problems, classification module is added in the pre-training process for the encoder to learn some domain specific information. We elaborate these three parts in the following.

Feature Disentangling. In order to take full advantage of the encoder, we adopt a two-branches structure encoder rather than splitting middle feature into two parts to get expression-related feature and identity-related feature. As figure 3 suggests, EIDNet encodes the neutral facial image I_N and the expression facial image I_E by the encoder and gets their embeddings, $[f_N^e, f_N^i]$ and $[f_E^e, f_E^i]$, separately. To our understanding, neutral facial image only embraces identity information while expression facial images embraces both identity information and expression information. Since these two images are from the same identity, their identity-related feature, i.e., f_N^i and f_E^i should be similar. The similarity loss is introduced as:

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

Expression Reconstruction. We believe that facial expression is shown on the face by comparing to the neutral state, i.e., non-expression state. Thus expression-related features can be extracted from expression facial image by comparing to neutral facial image. But the encoder structures of expression-related and expression-related features are the same, we need some constraints to obtain our goal. One possible method is forcing the expression-related features of neutral facial images to be zero. However, since the expressions of micro expression images are very subtle, the expression-related feature of expression facial image $[f_E^e$ itself has small intensities. Thus it will result in poor performance. In our method, we use reconstruction module instead. We believe that if we can reconstruct the original expression facial image I_E from concatenating f_N^i and f_E^e, i.e., $[f_E^e, f_N^i]$, the expression-related feature f_E^e actually conveys enough expression information of I_E. We introduce a decoder in the EIDNet to accomplish reconstruction. The expression reconstruction loss is:

$$\mathcal{L}_{rec} = \| I_E - decoder([f_E^e, f_N^i]) \|_2 \tag{2}$$

Classification. Many feature disentangle networks do not contain classification modules. They just use feature disentangle as a self-supervised learning tool and train a classifier in later part. However, we add classification in our methods mainly for two reasons. Firstly, expression-related features are really hard to be disentangled from identity-related features only by self-supervised methods, especially for micro-expression images. Secondly, since the goal our phase I is to produce pre-trained networks for adversarial learning in phase II, adopting classification module can help to force features more suited for their own domains, i.e., micro-expression and facial expression recognition problems. We conduct classification in EIDNet by adding several fully-connected layers after the encoder branch producing expression-related features. Normal Cross Entropy loss is used:

$$\mathcal{L}_{cls1} = CrossEntropy\{y_E - classifier(f_E^e)\} \tag{3}$$

where y_E softmax version of the ground truth label of I_E.

Thus the overall loss of phase I is:

$$\mathcal{L}_{phaseI} = \mathcal{L}_{cls1} + \lambda_{1,1}\mathcal{L}_{rec} + \lambda_{1,2}\mathcal{L}_{sim} \tag{4}$$

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

4.2. Adversarial Learning Framework

Assume two EIDNets already been pre-trained by micro-expression and macro-expression databases separately as described in the above section, named MicroNet and MacroNet. We will then conduct adversarial learning between features of micro-expression images and macro-expression images with same labels derived from MicroNet and MacroNet separately. Due to the reason that not all micro-expression images have corresponding same labeled macro-expression images, triplet term is added to solve this problem. Classification part is again added to control recognition accuracy of MicroNet and a new LIR term is introduced to guide classification loss between MicroNet and MacroNet. Triplet inputs will be used, which consist a micro-expression anchor I_{anc}, a same label macro-expression positive I_{pos} and a different label micro-expression negative I_{neg}. We will elaborate these modules below.

Triplet term. MacroNet is fixed while MicroNet will be further trained. We actually want to get useful information from MacroNet to MicroNet for performance boosting of micro-expression recognition by leverage macro-expression databases. For every triplet, the anchor and negative will be passed through the MicroNet and the positve will be passed through the MacroNet. Three corresponding expression embeddings can be get: f_{anc}^e, f_{pos}^e and f_{neg}^e and a triplet loss is introduced at the feature level:

$$\mathcal{L}_{tri} = max\{\| f_{anc}^e - f_{pos}^e \|_2 - \| f_{anc}^e - f_{neg}^e \|_2 + m, 0\} \tag{5}$$

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

Adversarial learning module. An adversarial learning protocol is added between MicroNet and MacroNet. Since micro-expression anchor and macro-expression positive in one triplet has the same label, we hope that by adopting adversarial learning, their expression embeddings can show similar distributions. Following this spirit, we adopt principles in [6]. The fixed MacroNet offers expression embedding of macro-expression images, tagged true labels; while MicroNet acts as the generator to give expression embeddings of micro-expression images, tagged false labels. A discriminator is introduced to identify these two labeled embeddings. Our MicroNet aims to generate micro-expression embeddings that the discriminator can not distinguish from same expression labeled macro-expression embeddings; while the discriminator aims to distinguish between these two kinds of embeddings. By this way, MicroNet can be fine tuned to embed micro-expression images of similar features with that of same labeled macro-expression images. The objective of our adversarial learning is thus as:

$$\min_{\theta_{Mi}} \max_{\theta_D} \{\mathbb{E}_{f^e_{pos} \sim P(f^e_{pos})} \log D_{\theta_D}(f^e_{pos}) + \mathbb{E}_{f^e_{anc} \sim P(f^e_{anc})} \log D_{\theta_D}(1 - f^e_{anc})\} \tag{6}$$

As shown in [6], the above equation can not been optimized directly thus new kind of loss is defined as:

$$\mathcal{L} = -\log D_{\theta_D}(f^e_{pos}) - \log D_{\theta_D}(1 - f^e_{anc})$$

Follow the suggestion in [6], it is better to minimize $-\log D_{\theta_D}(f^e_{anc})$ instead of minimizing $-\log D_{\theta_D}(1 - f^e_{anc})$ in order to avoid flat gradients. The adversarial loss is defined as:

$$\mathcal{L} = -\log D_{\theta_D}(f^e_{anc}) \tag{7}$$

where D is the discriminator, θ_D and θ_{Mi} are the parameters of the discriminator and the MicroNet, i.e., generator, separately.

Classification terms. The same classification loss is used to control recognition accuracy. Thus again, the cross entropy loss in added:

$$\mathcal{L}_{cls2} = CrossEntropy\{y_{anc} - classifier(f^e_{anc})\} \tag{8}$$

where y_{anc} is the softmax version of the ground truth label of I_{anc}.

Since we consider macro-expression databases as privileged information for our micro-expression recognition, we add Loss Inequality Regularization (LIR) [33]. The condition of using LIR loss is that privileged information is more discriminative than the primary features. In our understanding, macro-expressions are more separable than micro-expressions, thus it is reasonable to use LIR loss:

$$\mathcal{L}_{LIR} = max\{\mathcal{L}_{cls} - \mathcal{L}_{cls2}, 0\} \tag{9}$$

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 point y_{pos}.

Thus the overall loss of phase II is:

$$\mathcal{L}_{phaseII} = \mathcal{L}_{cls2} + \lambda_{2,1}\mathcal{L}_{tri} + \lambda_{2,2}\mathcal{L}_{adv} + \lambda_{2,3}\mathcal{L}_{LIR} \tag{10}$$

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

5. Experiment Results

5.1. Implementation Details

Structures of Encoder, Decoder and Discriminator. Structures of the encoder, decoder and discriminator are shown in Figure 3. ResNet18 is chosen 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 same branches linked after the backbone will give expression and identity embeddings separately. The backbone is shared since early features are general. For decoder, it uses up-sampling to double feature map size and also implements convolutional layers with ReLU and Batch normalization for information learning. The discriminator is responsible for identify true or false of input features. The structure is that several convolutional layers ending with a linear layer outputs a scalar value.

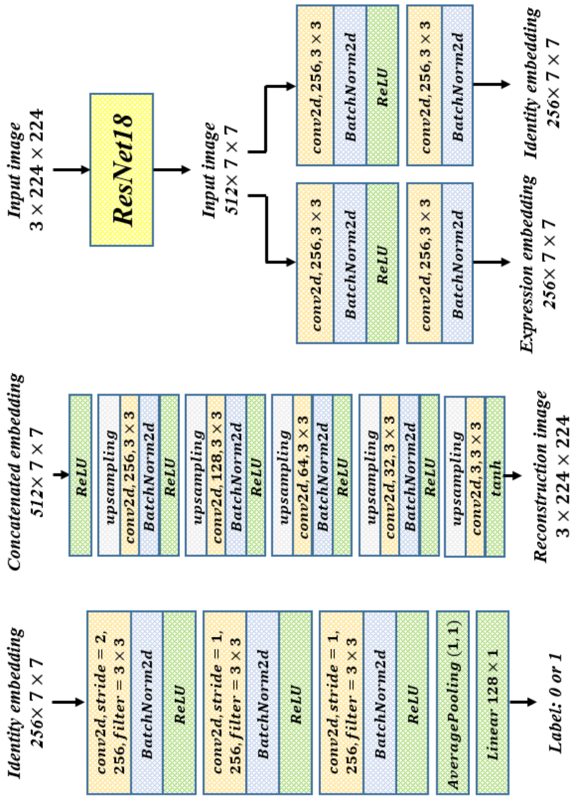

Figure 3. Figure of encoder, decoder and discriminator. With no special annotations, padding=1 and stride=1 for all conv2d modules.

Training Methods. As elaborations in section 3, there are six losses introduced by our complicated model thus training the whole framework in one step is infeasible. Lots of efforts of the optimizer will be devoted to influential losses such as L_cls and subtle losses like L_adv will be neglected. Thus we propose a two-phases training protocol, while in the first phase, we will pre-train two EIDNets with micro and macro expression databases separately, named MicroNet and MacroNet. While adversarial learning will be on the stage in second training phase for fine tuning. We elaborate these two phases in the following.

For training phase I, paired inputs including a neutral image and an expression image derived from the same video clip are used. Classification loss L_{cls}, similarity loss L_{sim} and reconstruction loss L_{rec} are used in phase I training for good feature disentanglement effect. After training, we believe that our MicroNet and MacroNet are capable to disentangle identity and expression embeddings, while images with different expressions can also be roughly separated. The whole process of training phase I is shown in figure **??**.

Pre-trained MicroNet and MacroNet are used in training phase II with MacroNet fixed. Triplet loss L_{tri}, adversarial loss L_{adv}, LIR loss L_{LIR} and again classification loss L_{cls} are implemented in this phase. In order to implement these losses, triplet inputs, including a pair of micro-expression image (anchor) and macro-expression image (positive) with the same label and another micro-expression image with different label (negative), are used. For each triplet input, triplet loss are obtained between all expression embeddings of three images, while classification loss only involve anchor; adversarial and LIR losses are between anchor and positive. We believe in this way, MicroNet can learn useful information from MacroNet and thus boosting the performance of micro-expression recognition. Figure **??** shows the training phase II process.

We set $\lambda_{1,1} = \lambda_{1,2} = 0.1$ in training phase I. Starting learning rates of the encoder, classifier and decoder are set to 10^{-4}, 10^{-4} and 10^{-5} separately. Every fold of LOSO procedure is trained with total 30 epochs.

For training phase II, $\lambda_{2,1} = \lambda_{2,2} = \lambda_{2,3} = 10^{-3}$ and the starting learning rates of the generator, discriminator and the classifier are set to 5×10^{-6}, 10^{-5} and 5×10^{-6}. The total epoch of each fold of LOSO procedure is 30. CosineAnnealingLR [21] is adopted in two phases to adjust learning rate with parameter $T_{max} = 15$.

Databases. As the micro-expression community always use CASME2 [34], SMIC [14] and SAMM [2] databases as evaluation standards for recognition tasks [23, 29], we adopt this custom in our paper. We mainly conduct two expriments on these databases. The first experiment is that we test our proposed framework on CASME2 and SMIC databases separately. The other experiment is that we adopt our framework on Composite Database Evaluation (CDE) task [23, 29], i.e. samples from those three databases are combined into a single composite database based on the reduced emotion classes. Leave-one-subject-out (LOSO) cross-validation is used to determine the training-test splits in both two experiments.

We use CK+ database [22] for the macro-expression database in the adversarial learning module as in many papers [10, 26, 1, 19]. CK+ database contains basic seven facial expressions plus neutral, while CASME2 database consists of data points with those labels: disgust, repression, happiness, surprise, sadness, fear and others. Since the number of data points with fear and sadness labels are too few, five-labels recognition tasks are normally implemented on CASME2 database with fear and sadness discarded [17, 19, 36, 31]. SMIC database categories all micro-expression into three labels: negative, positive and surprise. SAMM is a quite new database on which there are few experiments done. Since we need to use triplet inputs, which contain images with the same macro and micro expressions, in our training phase II, only related part of CK+ database are used in each experiment corresponding to each micro-expression database.

We do not use all frames in micro-expression databases, since many frames contains little or none additional information to neutral frame 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 (for situations where there are not enough frames after apex ones, choose more frames before as supplements).

Evaluation Protocols. Different evaluation protocols are adopted in our two experiments in order to compare with other methods. For the experiment that testing on single database, CASME2 and SMIC, normal accuracy and F1 score are used for evaluations. For the CDE task, we will use unweighted F1 score (UF1) and unweighted accuracy (UAR) introduced in [29].

5.2. Results and Analysis

We test our framework on two popular micro-expression tasks: 1) recognitions on CASME2 and SMIC databases separately and 2) composite database evaluation (CDE) task on combined CASME2, SMIC and SAMM databases. Comparisons with other state-of-the-art methods show the superiority of our proposed protocol.

Recognition on CASME2 and SMIC databases. We compare our framework with other eleven methods on CASME2 and SMIC databases. These methods are: 1) LBP-TOP [12], LBP-SIP [32], STLBP-IP [8], STCLQP [9], which are LBP based methods 2) HIGO [13], FHOFO [7], MDMO [20], Bi-WOOF [18], which are optimal flow based methods, and 3) Only-Apex [15], OFF-Apex [5], CNN+LSTM [11], which are deep feature methods. In the meantime, recognition results of only phase I and phase I + phase II of our framework are also offered to show the effects of adversarial learning. Table 1 shows the recognition results including accuracy and F1 score. We can draw the following observations.

Firstly, We can see that our framework after two phases training exceed other methods in almost every evaluation indicators on both databases. From the table, we can see deep feature methods are better than most handcraft feature methods, i.e., LBP and optimal flow based methods. For CASME2 database, our framework achieves a nearly 6% increase in accuracy and a 3% increase in F1 score com-

	Method	CASME2		SMIC	
		Accuracy	F1 Score	Accuracy	F1 Score
Handcraft Feature	LBP-TOP [12]	0.4900	0.5100	0.5800	0.6000
	LBP-SIP [32]	0.4656	0.4480	0.4451	0.4492
	STLBP-IP [8]	0.5951	0.5700	0.5793	0.5800
	HIGO [13]	0.6721	-	0.6829	-
	FHOFO [7]	0.5664	0.5248	0.5183	0.5243
	STCLQP [9]	0.6402	0.6381	0.5839	0.5836
	MDMO [20]	0.6737	-	**0.8000**	-
Deep Feature	Only-Apex [15]	0.6330	-	-	-
	Bi-WOOF [18]	0.5885	0.6100	0.6220	0.6200
	OFF-Apex [5]	-	-	0.6768	0.6709
	CNN+LSTM [11]	0.6098	-	-	-
Our Method	Phase I	0.6506	0.5501	0.6524	0.6465
	Phase I + Phase II	**0.7309**	**0.6640**	0.7317	**0.7249**

Table 1. Accuracy and F1 Score results on CASME2 and SMIC databases separately.

Method	Full		SMIC		CASME2		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [35]	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
OFF-Apex [5]	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
Bi-WOOF [18]	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
Neural [19]	0.7885	0.7824	0.7461	0.7530	0.8293	0.8209	0.7754	0.7152
Shallow [17]	0.7353	0.7605	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810
Dual [36]	0.7322	0.7278	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663
Capsule [31]	0.6520	0.6506	0.5820	0.5877	0.7068	0.7018	0.6209	0.5989
Phase I	0.6768	0.6809	0.5715	0.6119	0.7489	0.7271	0.6080	0.5671
Phase I + Phase II	**0.8647**	**0.8559**	**0.8582**	**0.8571**	**0.8863**	**0.8795**	**0.8135**	**0.8571**

Table 2. UF1 and UAR results of CDE task.

pared to the best results of previous methods. And for SMIC database, accuracy of our method is only below one method [20] and higher than other ones, while the F1 score of our method exceeds all others.

Secondly, By comparing results of our framework on phase I only and on phase I + phase II, we can show that our adversarial learning protocol really gives useful help in boosting performance of micro-expression recognition. We gain an 8% and an 9% increases in accuracy and F1 score on CASME2 database; and both 8% increases in accuracy and F1 score on SMIC database, by adopting phase II fine tuning on the basis of phase I training.

Finally, since data points in both CASME2 and SMIC databases are highly inbalanced, i.e., the numbers of data points with different labels are highly different, F1 score is more persuasive than accuracy. We can see that F1 scores of our framework after two phases training on both databases achieve the highest, showing the superiority of our framework.

Experiments of CDE task. CDE task is proposed by [23, 29] to evaluate micro-expression recognition on composite databases including CASME2, SMIC and SAMM. Tasks on cross databases have both pros and cons. The valuable training data points will be more sufficient for training and this is essential for micro-expression recognition since the available data points are too scarce. But since subjects in different databases are different, training across different databases will suffer from identity-related features distortion. However, our proposed framework adopts EIDNet as feature extractor can avoid this disadvantage and take full use of this composite databases. Extensive experiments confirm our analysis.

From table 2, we can see that the UF1 and UAR scores of our method after two-phases training all exceed 80% on full or single databases. And again, we compared our framework with other state-of-the-art methods including 1) LBP-TOP [35], BI-WOOF [18], which are handcraft feature methods, and 2) OFF-Apex [5], Neural [19], Shallow [17], Dual [36] and Capsule [31], which are deep feature method. We can get the following conclusions from complete results

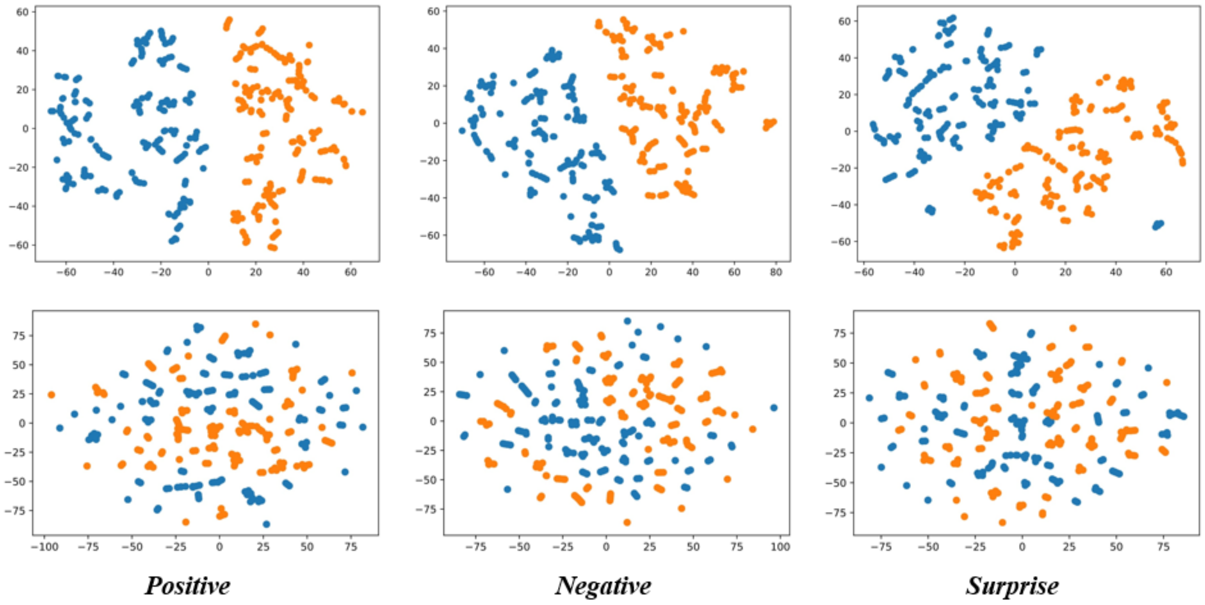

Figure 4. Figure of t-SNE results. The purple points in each image of first row correspond to dimensionality reduced micro-expression features of model trained after phase I and the second row corresponds to those after two-phases training, while the orange points are the same macro-expression features. Each column represent result of one expression labels.

of Table 2.

Firstly, deep feature methods all gain higher UF1 and UAR compared to handcraft feature methods. Actually, LBP-TOP [35], BI-WOOF [18] and OFF-Apex [5] are used as baseline in [29] for competition and Neural [19] wins the first prize in 2019. However, the results of our proposed framework exceeds the results of [19] in full or separated single databases of both UF1 and UAR: an $7.6\%/7.3\%$ increase in full database, an $11.2\%/10.4\%$ increase in SMIC database, an $5.7\%/5.7\%$ increase in CASME2 database and an $3.8\%/14.2\%$ increase in SAMM database of UF1/UAR.

Secondly, compared with results of our framework after only phase I training, the results of our framework after two-phases training gain incredible increases, with $18.8\%/17.5\%$ in full database, $28.7\%/16.6\%$ in SMIC database, $13.8\%/15.2\%$ in CAMSE2 database and $20.5\%/29.0\%$ in SAMM database of UF1/UAR. We can see the boosting effects of training phase II are very dramatic and this confirms the necessity and effectiveness of adversarial learning.

5.3. Visualization results

Since there are still four losses in phase II training prcocess, including $\mathcal{L}_{adv}$, $\mathcal{L}_{tri}$, $\mathcal{L}_{cls}$ and $\mathcal{L}_{LIR}$, we can not make sure that the increase of performance should attribute to adversarial learning. Thus we conduct t-SNE visualization methods to show that the feature distributions similarity between micro-expression and macro-expression images with the same label truly increase after adopting adversarial learning.

We conduct t-SNE analysis of features extracted from model trained in CDE task with three labels, positive, negative and surprise of only phase I training and two-phases training. For each expression label, randomly 200 samples are chosen from micro composite database and CK+ databases separately. From Figure 4 we can see that in all three expression label situations, features from micro-expression and macro-expression databases separate well from each others only after training phase I and correlate with each other after two-phases training, which verifies our hypothesis that adversarial learning will force micro-expression features being similar to same labeled macro-expression features.

6. Conclusion

Our paper presented an adversarial micro-expression recognition system by leveraging macro-expression databases as privileged information. An Expression-Identity Disentangle Network is also proposed to extract expression embeddings from original expression image input without identity-related information. Extensive experiments 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 also cast a light into a research direction of combining micro and macro expression recognition problems.

References

[1] Xianye Ben, Xitong Jia, Rui Yan, Xin Zhang, and Weixiao Meng. Learning effective binary descriptors for micro-expression recognition transferred by macro-information. *Pattern Recognition Letters*, 107:50–58, 2018. 2, 6

[2] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing*, 9(1):116–129, 2016. 6

[3] Paul Ekman. Lie catching and microexpressions. *The philosophy of deception*, pages 118–133, 2009. 1

[4] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009. 1

[5] YS Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. Off-apexnet on micro-expression recognition system. *Signal Processing: Image Communication*, 74:129–139, 2019. 1, 2, 6, 7, 8

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 4, 5

[7] SL Happy and Aurobinda Routray. Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Transactions on Affective Computing*, 2017. 1, 2, 6, 7

[8] Xiaohua Huang, Su-Jing Wang, Guoying Zhao, and Matti Piteikainen. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–9, 2015. 1, 2, 6, 7

[9] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng, and Matti Pietikäinen. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, 175:564–578, 2016. 1, 2, 6, 7

[10] Xitong Jia, Xianye Ben, Hui Yuan, Kidiyo Kpalma, and Weixiao Meng. Macro-to-micro transformation model for micro-expression recognition. *Journal of Computational Science*, 25:289–297, 2018. 2, 6

[11] Dae Hoe Kim, Wissam J Baddar, and Yong Man Ro. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 382–386. ACM, 2016. 1, 2, 6, 7

[12] Anh Cat Le Ngo, John See, and Raphael C-W Phan. Sparsity in dynamics of spontaneous subtle emotions: analysis and application. *IEEE Transactions on Affective Computing*, 8(3):396–411, 2016. 6, 7

[13] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE transactions on affective computing*, 9(4):563–577, 2017. 1, 2, 6, 7

[14] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*, pages 1–6. IEEE, 2013. 6

[15] Yante Li, Xiaohua Huang, and Guoying Zhao. Can micro-expression be recognized based on single apex frame? In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3094–3098. IEEE, 2018. 1, 2, 6, 7

[16] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019. 2

[17] Sze-Teng Liong, YS Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019. 1, 2, 6, 7

[18] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018. 1, 2, 6, 7, 8

[19] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4. IEEE, 2019. 1, 2, 6, 7, 8

[20] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing*, 7(4):299–310, 2015. 1, 2, 6, 7

[21] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[22] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010. 6

[23] Walied Merghani, Adrian Davison, and Moi Yap. 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)*, pages 662–666. IEEE, 2018. 6, 7

[24] Walied Merghani, Adrian K Davison, and Moi Hoon Yap. A review on facial micro-expressions analysis: datasets, features and metrics. *arXiv preprint arXiv:1805.02397*, 2018. 2

[25] Yee-Hui Oh, John See, Anh Cat Le Ngo, Raphael C-W Phan, and Vishnu M Baskaran. A survey of automatic facial micro-expression analysis: databases, methods, and challenges. *Frontiers in psychology*, 9:1128, 2018. 2

[26] Min Peng, Zhan Wu, Zhihao Zhang, and Tong Chen. From macro to micro expression recognition: deep learning on small datasets using transfer learning. In *2018 13th IEEE In-*

ternational Conference on Automatic Face & Gesture Recognition (FG 2018), pages 657–661. IEEE, 2018. 2, 6

[27] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, pages 808–822. Springer, 2012. 2

[28] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017. 1, 2

[29] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. Megc 2019–the second facial micro-expressions grand challenge. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019. 6, 7, 8

[30] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017. 2

[31] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. Capsulenet for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 1, 2, 6, 7

[32] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *Asian conference on computer vision*, pages 525–537. Springer, 2014. 1, 2, 6, 7

[33] Ziheng Wang and Qiang Ji. Classifier learning with hidden information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4969–4977, 2015. 5

[34] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014. 6

[35] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):915–928, 2007. 2, 7, 8

[36] Ling Zhou, Qirong Mao, and Luoyang Xue. Dual-inception network for cross-database micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019. 1, 2, 6, 7