Micro-Expression Recognition Enhanced by Macro-Expression from Spatial-Temporal Domain

Bin Xia[1] and **Shangfei Wang**[1,2] *

[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
xiabin@mail.ustc.edu.cn, sfwang@ustc.edu.cn

Abstract

Facial micro-expression recognition has attracted much attention due to its objectiveness to reveal the true emotion of a person. However, the limited micro-expression datasets have posed a great challenge to train a high performance micro-expression classifier. Since micro-expression and macro-expression share some similarities in both spatial and temporal facial behavior patterns, we propose a macro-to-micro transformation framework for micro-expression recognition. Specifically, we first pretrain two-stream baseline model from micro-expression data and macro-expression data respectively, named MiNet and MaNet. Then, we introduce two auxiliary tasks to align the spatial and temporal features learned from micro-expression data and macro-expression data. In spatial domain, we introduce a domain discriminator to align the features of MiNet and MaNet. In temporal domain, we introduce relation classifier to predict the correct relation for temporal features from MaNet and MiNet. Finally, we propose contrastive loss to encourage the MiNet to give closely aligned features to all entries from the same class in each instance. Experiments on three benchmark databases demonstrate the superiority of the proposed method.

1 Introduction

Micro-expression (ME) occurs when a person either deliberately or unconsciously conceals his or her genuine emotions [Ekman, 2009]. Compared to large intensity and long duration characteristics of macro-expression, micro-expressions is brief and subtle. Since it can be applied to many areas such as national security, clinical diagnosis and judicial system, automatic micro-expression recognition (MER) has become an active research area in recent years.

Micro-expression recognition can be classified into two categories: handcraft feature methods and deep feature methods. Researchers usually use Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOOF) and Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) to extract handcraft features. Li *et al.* [2017] used feature differences for ME spotting and adopted the histogram of image gradient orientation-TOP (HIGO-TOP) for MER. Liong *et al.* [2018] introduced a Bi-Weighted Oriented Optical Flow (Bi-WOOF) based feature extractor, while Happy *et al.* [2017] proposed a fuzzy HOOF method (FHOFO), which ignored the subtle motion magnitudes and only took the motion direction into consideration. Wang *et al.* [2014] adopted a pruned LBP descriptor using six neighbors around every point (LBP-SIP) which reduces the inherent redundancy within LBP-TOP. Huang *et al.* [2015] adopted an integral projection method to boost the capability of LBP-TOP (STLBP-IP) by supplementing shape information. Le *et al.* [2016] used LBP-TOP to learn significant temporal and spectral structures with sparsity constraints. Due to the short duration and low intensity of ME, handcraft features are not robust in the micro-expression identification and classification.

Recently researchers use deep network for MER as the micro-expression databases gradually developed. Peng *et al.* [2019] explored their underlying joint formulations and proposed a consolidated Eulerian framework to reveal the subtle facial movements. It expanded the temporal duration and amplified the muscle movements in ME simultaneously. Van Quang *et al.* [2019] used the newly proposed framework CapsuleNet [2017] to figure out the part-whole relationships for MER. Khor *et al.* [2019] proposed a lightweight dual-stream shallow network (DSSN) in the form of a pair of truncated CNNs with heterogeneous input features. Liong *et al.* [2019] designed a shallow triple stream three-dimensional CNN (STSTNet) to extract details of ME, with optical strain, horizontal and vertical optical flow as input. Lei *et al.* [2020] applied learning-based video motion magnification to magnify ME and designed graph-temporal convolutional network (Graph-TCN) to extract the features of the local muscle movements. To fully exploit the dependence between action units (AUs) and expression, Sun *et al.* [2020] proposed a knowledge transfer technique that distilled and transferred multi-knowledge from AU for MER. Xie *et al.* [2020] proposed AU-assisted graph attention convolutional network (AU-GACN) for MER, which effectively integrated AU recognition. Nevertheless, these deep learning methods suffer from insufficient micro-expression training samples.

Nowadays there are many macro-expression databases, which contain a large number of labeled training samples.

*Contact Author

Figure 1: We show some examples of micro-expression and macro-expression videos, where the first and second rows are happiness from SAMM and CK+ database respectively. The third and fourth rows are surprise from CASME II and MMI database. The expression videos change from onset to apex frame. The red box areas of apex frame reveal micro and macro expression share some similarity in facial behavior patterns.

Although macro-expression has longer duration and higher intensity than micro-expression, these two expressions share some similarities in spatial and temporal facial behavior patterns. Figure 1 shows a comparison between micro and macro expressions. As shown in the expression sequences varying from onset frame to apex frame, we can obviously find that both micro and macro expression gradually raise the lip corners for happiness. For surprise, both image sequences raise eyebrows and open eyes. Thus, how to transform macro-expression information for micro-expression recognition has become an important research direction.

Peng *et al.* [2018] pretrained a deep network from macro-expression images, and fine-tuned it with micro-expression images by transfer learning protocols. However, they did not consider the gap between micro and macro expression images, which limits the effect of transfer learning. Liu *et al.* [2019] used Expression Magnification and Reduction (EMR) to reduce the gap of apex frame between micro and macro expression visually, and trained on a fusion of micro and macro expression database (Neural). Since EMR actually introduces a lot of noise to micro-expression, it can't guarantee the similarity of micro and macro expression features. Xia *et al.* [2020] introduced disentangle network to extract expression-related embedding from apex frame, and used loss equality regularization to transform macro-expression information (MicroNet). However, their method can't be trained end-to-end, which impairs the performance of overall framework. The above macro-expression assisted methods only use spatial features to capture static structure patterns, which ignore dynamic muscle movement from temporal features.

In order to tackle the aforementioned challenges, in this paper we propose a macro-to-micro transformation model which enables to transfer spatial and temporal pattern existed in macro-expression to micro-expression recognition. As shown in Figure 2, we first pretrain two-stream baseline model for micro-expression data and macro-expression data respectively, named MiNet and MaNet. In order to take advantage of macro-expression data, we model the shared features in the spatial and temporal domain simultaneously. In spatial domain, we introduce a domain discriminator to align the features of MiNet and MaNet, so MiNet can capture static textures of facial appearances better. In temporal domain, we introduce a relation classifier to predict the correct relation for temporal features of different sampling interval from MaNet and MiNet. Through this task, MiNet can learn the dynamic pattern of muscles from MaNet. Finally, we explicitly take the class label into account and introduce contrastive loss to encourage the MiNet to pull the same class samples together in the feature space, while simultaneously pushing apart clusters of samples from different classes.

In summary, our contributions are two-folds: 1) We propose a well-designed macro-to-micro transformation framework by two auxiliary tasks from spatial and temporal domain respectively. 2) We utilize contrastive loss to learn a more robust clustering of the feature space for micro-expression and macro-expression.

2 Method

2.1 Problem Statement

Suppose we have micro-expression data $\mathcal{D}_I = \{x_i^{(j)}, y_i^{(j)}\}_{j=1}^{N}$, and macro-expression data $\mathcal{D}_A = \{x_a^{(j)}, y_a^{(j)}\}_{j=1}^{M}$. $\mathcal{D}_I$ contains N training instances of micro-expression and $\mathcal{D}_A$ contains M training instances of macro-expression. x_i and x_a respectively represent the micro and macro expression videos that change from neutral-face to emotional-face. $y_i, y_a \in \{1, 2, \cdots, L\}$ are the expression label, L is the number of expression category. Our goal is to train a deep network for micro-expression recognition with the help of macro-expression data in both spatial and temporal domain. During testing, only micro-expression data is required.

2.2 Baseline Model

Videos can be decomposed into spatial and temporal components. The spatial part, in the form of individual frame appearance, carries static structure patterns about face depicted in the video. The temporal part, in the form of motion across the frames, conveys the movement of the facial muscle.

We use two-stream network as our baseline model, which contains spatial stream network, temporal stream network and expression classifier. We pick out the apex frame from the expression video and input it to the spatial stream network. The spatial stream network operates on individual image frame, effectively extracting apex-level feature f^A from apex frame x^A. The appearance from apex frame contains useful clue, since some facial expressions are strongly associated with particular facial muscle. The temporal stream network is proposed to capture dynamic patterns from consecutive frames. The temporal stream network first extracts frame-level features $f = \{f^1, f^2, \cdots, f^T\}$ from expression video, and T is the length of video. Then we aggregate frame-level features with mean-pooling along the temporal direction

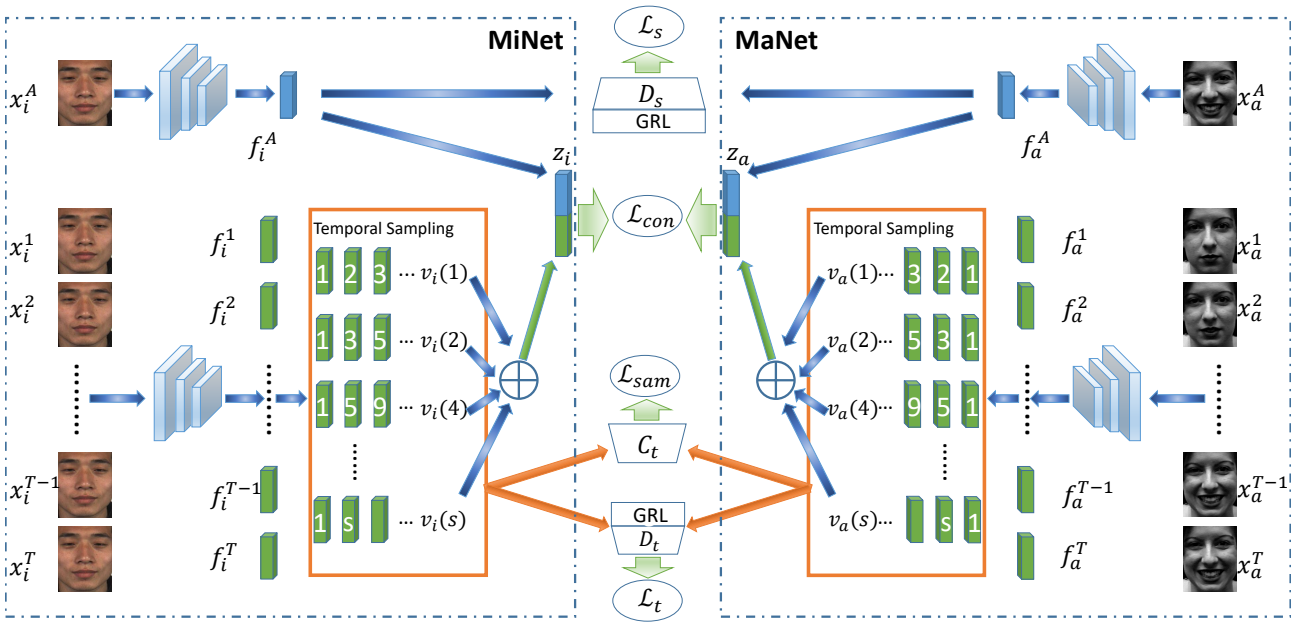

Figure 2: The framework of our micro-expression recognition model. First we pretrain two-stream with micro and macro expression databases separately, named MiNet and MaNet. Secondly, MaNet is used to guide the fine-tuning of MiNet from both spatial and temporal domain.

to generate the video-level feature v. We fuse the apex-level feature f^A and video-level feature v to get fused feature z, and input it into expression classifier C_e.

Two two-stream networks are built in our framework. The first one performs micro-expression recognition from micro-expression data $\mathcal{D}_I$, and is the desired classifier in this paper, named MiNet. The second one performs macro-expression recognition from macro-expression data $\mathcal{D}_A$, named MaNet. These two networks are pretrained with the supervised multi-class cross entropy loss as shown in Eq.1:

$$\mathcal{L}_c = -\sum_{l=1}^{L} 1_{[y=l]} \log(C_e(z)) \tag{1}$$

$1_B \in \{0, 1\}$ is an indicator function that returns 1 if B evaluates as true. After pretraining, MiNet is further fine-tuned under the guidance of MaNet.

2.3 Guidance from Spatial Domain

The spatial stream network can process the static facial textures, which have a strong correlation with expression. Limited by small datasets of micro-expression, the spatial stream of MiNet may not capture static textures of facial appearances well. Thus, we propose an auxiliary task to learn effective feature from macro-expression in spatial domain.

Inspired by one of the most popular adversarial-based approaches DANN [Ganin and Lempitsky, 2015], we adopt adversarial method to align the spatial features of MiNet and MaNet. Specifically, for the spatial features from the MiNet and MaNet, we introduce spatial domain discriminator D_s to align macro-expression features f_a^A and micro-expression features f_i^A. Through adversarial training with a gradient reversal layer (GRL) which reverses the gradient signs during back-propagation, MiNet is optimized to gradually align the feature distributions in the spatial domain of micro-expression to that of macro-expression. Given a binary domain label d, indicating if a feature $f^A \in f_a^A$ or $f^A \in f_i^A$, the outputs of D_s are used to calculate the spatial domain loss, which can be defined as:

$$\mathcal{L}_s = -d \log(D_s(f^A)) - (1-d) \log(1 - D_s(f^A)) \tag{2}$$

2.4 Guidance from Temporal Domain

Although spatial stream network can effectively derive spatial patterns from still images, it cannot capture the temporal pattern in consecutive frames. The temporal pattern corresponds to the movement of facial muscles. Although we implicitly encode temporal features by the mean-pooling, the relation between frames is still missing. In order to address temporal variations for videos, we propose an auxiliary task for video-level features. This is a temporal relation classification task, which predicts the correct relation for temporal features of different sampling interval from both MaNet and MiNet.

Specifically, we first get frame-level features $f_i = \{f_i^1, f_i^2, \cdots, f_i^T\}$ from MiNet. Then, we uniformly sample a frame feature from each s frames with the same temporal interval to generate temporal features $v_i(s)$. The temporal features $v_i(s)$ with different sampling intervals have consistent context but different playback rates. We use different playback rates to learn slow and fast motion patterns simultaneously. Fast playrate can quickly understand video content, and slow playrate can capture fine details. Similarly, we extract macro-expression temporal features $v_a(s)$ from MaNet. We introduce relation classifier C_s to predict the sampling interval for $v_i(s)$ and $v_a(s)$. We set the sampling interval as

$s = 2^k (k = 0, 1, 2, \cdots, s_c - 1)$. The ground-truth label is denoted as $y_s \in \{1, 2, \cdots, s_c\}$, s_c is the number of different sampling intervals of the input features. We use cross entropy loss to supervise relation classification task as shown in Eq.3:

$$\mathcal{L}_{sam} = -\sum_{l=1}^{s_c} 1_{[y_s=l]} \log(C_s(v(s))) \tag{3}$$

where $v(s) \in \{v_i(s), v_a(s)\}$. The MiNet and MaNet are driven to perceive subtle difference among adjacent frames which is important to learn temporal pattern.

In order to align the temporal features, we also introduce temporal domain discriminator D_t to predict whether the temporal feature $v(s)$ is from micro-expression data or macro-expression data. Given a binary domain label d, indicating if a feature $v(s) \in v_i(s)$ or $v(s) \in v_a(s)$, the outputs of D_t are used to calculate the temporal domain adversarial loss, as shown in Eq.4:

$$\mathcal{L}_t = -d \log(D_t(v(s))) - (1 - d) \log(1 - D_t(v(s))) \tag{4}$$

Through adversarial training with GRL, $\mathcal{L}_t$ also contributes to optimize MiNet to align the feature distribution in the temporal domain of micro-expression to that of macro-expression.

Replacing original video-level feature v by aggregating multiple temporal features $v(s)(s = 1, 2, 4, \cdots, 2^{s_c-1})$, the MiNet can encode temporal relation information.

2.5 Contrastive Loss from Spatial-temporal Domain

Up to now, we only consider minimize the domain discrepancy between micro and macro expression, but neglect the class label, which may lead to misalignment and poor generalization performance. We propose to explicitly take the class label into account and measure the intra-class and inter-class discrepancy. The intra-class discrepancy is minimized to compact the features of micro and macro expression within a class, whereas the inter-class discrepancy is maximized to push the features of each class further away from the decision boundary. The contrastive objective functions [Khosla *et al.*, 2020] have achieved excellent performance in recent years by sampling positive pairs and negative pairs. Inspired by contrastive learning framework, for each anchor micro-expression sample, we generate many positive pairs and negative pairs from micro and macro expression minibatch.

Specifically, for a minibatch of n randomly sampled micro-expression samples $\{x_i^{(j)}, y_i^{(j)}\}_{j=1}^n$, the corresponding randomly sampled minibatch used for training consists of n macro-expression samples $\{x_a^{(j)}, y_a^{(j)}\}_{j=1}^n$. We input them into MiNet and MaNet respectively to get their fused features $\{z_i^{(j)}\}_{j=1}^n$, $\{z_a^{(j)}\}_{j=1}^n$. Within a minibatch, let $j \in \{1, \cdots, n\}$ be the index of an anchor sample, and generates positive pairs with the same expression labels from micro and macro expression data, generates negative pairs with remaining samples. The contrastive loss takes the following form:

$$\mathcal{L}_{con} = -\sum_{j=1}^{n} \mathcal{L}_{con}^j \tag{5}$$

Dataset \ Expression	Negative	Positive	Surprise	Total
SMIC	70	51	43	164
CASME II	88	32	25	145
SAMM	92	26	15	133
3DB-combined	250	109	83	442

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

$$\mathcal{L}_{con}^j = \sum_{k=1}^{n} 1_{[k \neq j]} 1_{[y_i^{(k)} = y_i^{(j)}]} \log \frac{e^{(sim(z_i^{(j)}, z_i^{(k)})/\tau)}}{D^j} + \sum_{k=1}^{n} 1_{[y_a^{(k)} = y_i^{(j)}]} \log \frac{e^{(sim(z_i^{(j)}, z_a^{(k)})/\tau)}}{D^j} \tag{6}$$

$$D^j = \sum_{u=1}^{n} e^{(sim(z_i^{(j)}, z_i^{(u)})/\tau)} + \sum_{u=1}^{n} e^{(sim(z_i^{(j)}, z_a^{(u)})/\tau)} \tag{7}$$

Let $sim(w, v)$ denote the cosine similarity between two vectors w and v, and τ denote a temperature parameter.

During training, for any anchor sample j, all positive pairs in a minibatch contribute to the numerator of Eq.6. This loss encourages the MiNet to give closely aligned features to all entries from the same class in each instance, resulting in a more robust clustering of the feature space for micro-expression and macro-expression.

2.6 Overall Loss Function and Optimization

Finally, by combining the losses $\mathcal{L}_c, \mathcal{L}_s, \mathcal{L}_{sam}, \mathcal{L}_t$ and $\mathcal{L}_{con}$, we define the general loss as follows:

$$\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_{sam} + \lambda_3 \mathcal{L}_t + \lambda_4 \mathcal{L}_{con} \tag{8}$$

where λ_1, λ_2, λ_3 and λ_4 are the hyperparameters controlling loss coefficients. We use the general loss to update the parameter of MiNet with Adam optimizer. After training, we use MiNet for micro-expression recognition.

3 Experiment

3.1 Experimental Condition

There are three databases commonly used for MER, i.e., CASME II [Yan *et al.*, 2014], SMIC [Li *et al.*, 2013] and SAMM [Davison *et al.*, 2016]. The CASME II [Yan *et al.*, 2014] has 256 micro-expression videos from 26 subjects, with the average age of 22.03 years old at 200 fps. The videos in this database show a participant evoked by one of five categories of micro-expressions: happiness, disgust, repression, surprise and others. The SMIC [Li *et al.*, 2013] database contains 164 micro-expression samples from 16 participants. Each micro-expression is recorded at the speed of 100fps and labeled with three general expression labels: positive, negative and surprise. The SAMM [Davison *et al.*, 2016] database 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. Seven micro-expression types are included in the SAMM dataset. They are happiness, surprise, disgust, repression, angry, fear and contempt.

Method	CASME II(CK+)		SMIC(CK+)		SAMM(CK+)		CASME II(MMI)		SMIC(MMI)		SAMM(MMI)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
$\mathcal{L}_c$	0.715	0.636	0.695	0.672	0.701	0.693	0.715	0.636	0.695	0.672	0.701	0.693
$\mathcal{L}_c + \mathcal{L}_s$	0.751	0.681	0.744	0.729	0.729	0.724	0.755	0.691	0.726	0.701	0.725	0.722
$\mathcal{L}_c + \mathcal{L}_{sam} + \mathcal{L}_t$	0.763	0.706	0.750	0.736	0.735	0.730	0.763	0.705	0.726	0.701	0.731	0.727
$\mathcal{L}_c + \mathcal{L}_s + \mathcal{L}_{sam} + \mathcal{L}_t$	0.771	0.722	0.762	0.750	0.751	0.748	0.779	0.733	0.762	0.750	0.741	0.740
$\mathcal{L}_c + \mathcal{L}_{con}$	0.747	0.683	0.738	0.738	0.731	0.726	0.755	0.694	0.726	0.703	0.725	0.722
$\mathcal{L}_c + \mathcal{L}_s + \mathcal{L}_{sam} + \mathcal{L}_t + \mathcal{L}_{con}$	**0.791**	**0.748**	**0.786**	**0.778**	**0.767**	**0.764**	**0.799**	**0.759**	**0.774**	**0.761**	**0.758**	**0.754**

Table 2: Accuracy and F1 Score results on the CASME II, 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.

Method	CK+		MMI	
	UF1	UAR	UF1	UAR
$\mathcal{L}_c$	0.801	0.798	0.801	0.798
$\mathcal{L}_c + \mathcal{L}_s$	0.843	0.840	0.835	0.835
$\mathcal{L}_c + \mathcal{L}_{sam} + \mathcal{L}_t$	0.855	0.853	0.851	0.848
$\mathcal{L}_c + \mathcal{L}_s + \mathcal{L}_{sam} + \mathcal{L}_t$	0.871	0.866	0.851	0.848
$\mathcal{L}_c + \mathcal{L}_{con}$	0.839	0.835	0.820	0.822
$\mathcal{L}_c + \mathcal{L}_s + \mathcal{L}_{sam} + \mathcal{L}_t + \mathcal{L}_{con}$	**0.883**	**0.876**	**0.863**	**0.860**

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

We conduct two experiments on these databases. First, we test our method on the CASME II, SMIC and SAMM databases separately. Accuracy and F1 score are used for evaluation. Second, we test our method on Composite Database Evaluation (CDE) task [See *et al.*, 2019], i.e., samples from all databases are combined into a single composite database based on the reduced expression classes. The distribution of samples and subjects are given in Table 1. Unweighted F1 score (UF1) and Unweighted Average Recall (UAR) are used for evaluation. Leave-one-subject-out (LOSO) cross-validation is used in all experiments.

Two popular lab-collected databases, i.e., CK+[Lucey *et al.*, 2010] and MMI[Pantic *et al.*, 2005] are adopted as macro-expression. The CK+ database contains 327 image sequences of seven expressions: anger, contempt, disgust, fear, happiness, sadness, and surprise. The MMI database includes 205 image sequences with frontal faces of six expression labels: anger, disgust, fear, happiness, sadness and surprise.

In our experiments, expression videos are preprocessed to a fixed length, and every frame is rescaled to a fixed size. We use ResNet18 as the architecture of spatial stream network and temporal stream network. The spatial discriminator and temporal discriminator are three convolutional layers ending with a linear layer outputs a scalar value. The structure of relation classifier is three convolutional layers ending with a linear layer and outputs the prediction of all possible sampling intervals. We use Eq.8 to update the parameter of MiNet with the Adam optimizer, and set $\lambda_1 = \lambda_4 = 0.1$, $\lambda_2 = \lambda_3 = 0.2$.

3.2 Experimental Results and Analysis

Recognition on the CASME II, SAMM and SMIC Databases Separately

We conduct ablation experiments to verify the influence of different loss functions, i.e., spatial domain loss $\mathcal{L}_s$, temporal domain loss $\mathcal{L}_{sam} + \mathcal{L}_t$ and contrastive loss $\mathcal{L}_{con}$ on the final recognition performance. As shown in Table 2, We can draw the following observations:

First, adopting one of the introduced losses leads to an improvement comparing with the baseline model only using $\mathcal{L}_c$. Specifically, the accuracy/F1 of $\mathcal{L}_c + \mathcal{L}_s$, $\mathcal{L}_c + \mathcal{L}_{sam} + \mathcal{L}_t$ and $\mathcal{L}_c + \mathcal{L}_{con}$ are 3.6%/4.5%, 4.8%/7.0% and 3.2%/4.7% higher than the baseline on the CASME II database, with CK+ database as macro-expression. The experimental results on the SAMM and SMIC databases show similar trend. Limited by the amount of micro-expression data, deep methods don't have good generalization. However, our method takes advantage of the macro-expression data effectively and transfers them to micro-expression recognition by adopting two auxiliary tasks in spatial domain and temporal domain and contrastive loss in spatial-temporal domain.

Second, the spatial and temporal domain losses achieve better performance than contrastive loss in most cases. For example, $\mathcal{L}_c + \mathcal{L}_s + \mathcal{L}_{sam} + \mathcal{L}_t$ gains 2.4%/3.9%, 2.4%/1.2% and 2.0%/2.2% in accuracy/F1 than $\mathcal{L}_c + \mathcal{L}_{con}$ on the CASME II, SMIC and SAMM database respectively, with CK+ database as macro-expression. Although contrastive loss takes the class label into account and measures the intra-class and inter-class discrepancy, it is difficult to choose a suitable distance metric.

Finally, our method combines the strengths of the four introduced loss functions and achieves the best performance. Specifically, the accuracy/F1 of our method is 7.6%/11.2%, 9.1%/10.6% and 6.6%/7.1% higher than the baseline on the CASME II, SAMM and SMIC database, with CK+ database as macro-expression. This indicates that the different guidance will not cause the inter-domain discrepancy. Two auxiliary tasks only align overall feature distribution and don't consider the distribution of each category, but contrastive loss takes the class information into account.

Recognition of CDE Task

CDE task is proposed to evaluate micro-expression recognition on composite database including CASME II, SMIC and SAMM. Tasks on cross databases have both pros and cons. The valuable training data will be more sufficient for training and this is essential for micro-expression recognition since the available data are too scarce. But since composite database comprises samples collected from different environment and subjects, training across different databases will suffer from domain shift problem. Since our method takes advantage of the macro-expression databases for micro-expression recognition, and adopts two auxiliary tasks to learn domain-invariant features, our method can avoid domain shift problem and take full use of composite

Method	CASME II		SMIC		SAMM	
	Acc	F1	Acc	F1	Acc	F1
LBP-TOP [Le Ngo *et al.*, 2016]	0.490	0.510	0.580	0.600	0.590	0.364
LBP-SIP [Wang *et al.*, 2014]	0.465	0.448	0.445	0.449	0.415	0.406
STLBP-IP [Huang *et al.*, 2015]	0.595	0.570	0.579	0.580	0.568	0.527
STCLQP [Huang *et al.*, 2016]	0.640	0.638	0.583	0.583	0.638	0.611
HIGO [Li *et al.*, 2017]	0.672	-	0.682	-	-	-
FHOFO [Happy and Routray, 2017]	0.566	0.524	0.518	0.524	-	-
Bi-WOOF [Liong *et al.*, 2018]	0.588	0.610	0.622	0.620	0.583	0.397
OFF-Apex [Gan *et al.*, 2019]	-	-	0.676	0.670	0.681	0.542
Boost [Peng *et al.*, 2019]	0.709	-	0.689	-	-	-
DSSN [Khor *et al.*, 2019]	0.708	0.730	0.634	0.646	0.574	0.464
AU-GACN [Xie *et al.*, 2020]	0.712	0.355	-	-	0.702	0.433
Dynamic [Sun *et al.*, 2020]	0.726	0.670	0.761	0.710	-	-
MicroNet [Xia *et al.*, 2020]	0.756	0.701	0.768	0.744	0.741	0.736
Graph-TCN [Lei *et al.*, 2020]	0.740	0.725	-	-	0.750	0.699
ours	**0.799**	**0.759**	**0.786**	**0.778**	**0.767**	**0.764**

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

Method	Full		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [Zhao and Pietikainen, 2007]	0.588	0.578	0.200	0.528	0.702	0.742	0.395	0.410
Bi-WOOF [Liong *et al.*, 2018]	0.629	0.622	0.572	0.582	0.780	0.802	0.521	0.513
OFF-Apex [Gan *et al.*, 2019]	0.719	0.709	0.681	0.669	0.876	0.868	0.540	0.539
Capsule [Van Quang *et al.*, 2019]	0.652	0.650	0.582	0.587	0.706	0.701	0.620	0.598
Shallow [Liong *et al.*, 2019]	0.735	0.760	0.680	0.701	0.838	0.868	0.658	0.681
Dual [Zhou *et al.*, 2019]	0.732	0.727	0.664	0.672	0.862	0.856	0.586	0.566
Neural [Liu *et al.*, 2019]	0.788	0.782	0.746	0.753	0.829	0.820	0.775	0.715
MicroNet [Xia *et al.*, 2020]	0.864	0.857	0.864	0.861	0.870	0.872	0.825	0.819
ours	**0.883**	**0.876**	**0.873**	**0.867**	**0.881**	**0.881**	**0.896**	**0.884**

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

database. As shown in Table 3, the UF1/UAR of our method is 8.2%/7.8% and 6.2%/6.2% higher than the baseline on the CDE task, with CK+ and MMI as macro-expression dataset.

3.3 Comparison with Related Works

We compare our framework with related works. These methods are: 1) LBP-TOP [2016], LBP-SIP [2014], STLBP-IP [2015], STCLQP [2016], HIGO [2017], FHOFO [2017] and Bi-WOOF [2018], which are handcraft feature methods, 2) OFF-Apex [2019], Boost [2019], DSSN [2019], AU-GACN [2020], Dynamic [2020], Graph-TCN [2020], Shallow [2019], Dual [2019] and Capsule [2019], which are deep feature methods, 3) Neural [2019] and MicroNet [2020], which are macro-expression assisted methods.

From Table 4, we can see that our framework exceeds most handcraft feature methods in almost every evaluation indicators. Our framework achieves nearly 21.1%, 16.4% and 18.4% increases in accuracy, 14.9%, 15.8% and 36.7% increases in F1 score compared to the best results of handcraft feature methods, i.e., Bi-WOOF on the CASME II, SMIC and SAMM database. Due to short duration and low intensity of micro-expression, these handcraft feature methods can't capture the details of facial appearance. Our method also outperforms the best results of deep feature methods, i.e., Graph-TCN by $5.9\%/3.4\%$ and $2.7\%/6.5\%$ on the CASME II and SAMM database of accuracy/F1. Graph-TCN applied video motion magnification to magnify the intensity of micro-expression, but this operation would introduce noise to micro-expression data. In this paper we introduce auxiliary tasks to make micro and macro expression samples produce similar feature distributions in spatial and temporal domain.

As shown in Table 5, our method gains higher results on CDE task. Because the CDE task has greatly increased micro-expression training data, deep feature methods (i.e., Shallow, Dual and Capsule) can't handle the difference between databases very well. As a result, these methods perform well on the CASME II database, but get poor results on the SMIC and SAMM databases. However, our proposed framework adopts two auxiliary tasks to learn domain-invariant features from macro-expression data, which avoid domain shift problem and take full use of composite database.

Compared with the macro-expression assisted methods, i.e., Neural and MicroNet, our proposed method exceeds their results in both composite and single databases. Neural used EMR to reduce the gap between micro and macro expression visually. This preprocessing 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 features. MicroNet can't be trained end-to-end, which impairs the performance of overall framework. Neural and MicroNet only capture static structure patterns from spatial domain, and ignore temporal pat-

terns. However, our method adopts a temporal relation classification task to learn dynamic movement patterns.

4 Conclusion

In this paper we propose a macro-to-micro transformation model which transfers spatial and temporal pattern existed in macro-expression to micro-expression recognition. In order to take advantage of macro-expression data, we introduce adversarial learning to align the spatial features of MiNet and MaNet. In temporal domain, our proposed auxiliary task predicts temporal relation. Through this task, MiNet can learn the dynamic movement relationship of the muscles from MaNet. Finally, the proposed contrastive loss encourages the MiNet to give closely aligned features to all entries from the same class in each instance. Experiments on three benchmark databases demonstrate that our framework outperforms the state-of-the-art micro-expression recognition methods.

Acknowledgments

This work was supported by the National Key R & D program of China 2020YFC2007700, project from Anhui Science and Technology Agency 1804a09020038.

References

[Davison *et al.*, 2016] 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.

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

[Gan *et al.*, 2019] 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.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

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

[Huang *et al.*, 2015] 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.

[Huang *et al.*, 2016] 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.

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

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[Le Ngo *et al.*, 2016] 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.

[Lei *et al.*, 2020] Ling Lei, Jianfeng Li, Tong Chen, and Shigang Li. A novel graph-tcn with a graph structured representation for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2237–2245, 2020.

[Li *et al.*, 2013] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *FG 2013*, pages 1–6. IEEE, 2013.

[Li *et al.*, 2017] 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.

[Liong *et al.*, 2018] 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.

[Liong *et al.*, 2019] 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 *FG 2019*, pages 1–5. IEEE, 2019.

[Liu *et al.*, 2019] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *FG 2019*, pages 1–4. IEEE, 2019.

[Lucey *et al.*, 2010] 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.

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

[Peng *et al.*, 2018] Min Peng, Zhan Wu, Zhihao Zhang, and Tong Chen. From macro to micro expression recognition: deep learning on small datasets using transfer learning. In *FG 2018*, pages 657–661. IEEE, 2018.

[Peng *et al.*, 2019] Wei Peng, Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. A boost in revealing subtle facial expressions: A consolidated eulerian framework. In *FG 2019*, pages 1–5. IEEE, 2019.

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

[See *et al.*, 2019] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. Megc 2019–the second facial micro-expressions grand challenge. In *FG 2019*, pages 1–5. IEEE, 2019.

[Sun *et al.*, 2020] Bo Sun, Siming Cao, Dongliang Li, Jun He, and Lejun Yu. Dynamic micro-expression recognition using knowledge distillation. *IEEE Transactions on Affective Computing*, 2020.

[Van Quang *et al.*, 2019] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. Capsulenet for micro-expression recognition. In *FG 2019*, pages 1–7. IEEE, 2019.

[Wang *et al.*, 2014] 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.

[Xia *et al.*, 2020] Bin Xia, Weikang Wang, Shangfei Wang, and Enhong Chen. Learning from macro-expression: a micro-expression recognition framework. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2936–2944, 2020.

[Xie *et al.*, 2020] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. Au-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2871–2880, 2020.

[Yan *et al.*, 2014] 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.

[Zhao and Pietikainen, 2007] 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.

[Zhou *et al.*, 2019] Ling Zhou, Qirong Mao, and Luoyang Xue. Dual-inception network for cross-database micro-expression recognition. In *FG 2019*, pages 1–5. IEEE, 2019.