Compound Expression Recognition In-the-wild with AU-assisted Meta Multi-task Learning

Ximan Li[1], Weihong Deng[1]*, Shan Li[1], Yong Li[2]
[1] Beijing University of Posts and Telecommunications, Beijing,China
[2] Nanjing University of Science and Technology, Nanjing, China
{liximan, whdeng and ls1995}@bupt.edu.cn, yong.li@njust.edu.cn

Abstract

Facial expression recognition (FER) has received wide attention as an essential part of affective computing. Considering its ambiguity and variety, more attention has been paid to compound expression recognition. Since emotions are generated by the contraction of muscle groups, the action units (AUs) analysis has a vital role in FER. However, AU analysis of compound expression has only been conducted in the laboratory, lacking real-world databases with manually annotated compound expressions and AUs. We construct a real-world affective faces database of compound emotions (RAF-CE), with both compound expression labels and AU labels. Our AU analysis of compound facial expressions conducted on RAF-CE reveals that AU patterns and AU frequencies are different in the lab-controlled compound expressions and the real-world ones. Based on the analysis, we propose a meta-based multi-task learning (MML) for compound FER with AU recognition utilized as an auxiliary task. To fully exploit the priori AU-emotion constraint observed in RAF-CE, an alignment loss is introduced to explicitly match the distribution of AU and FE predictions with each other. Furthermore, we adopt meta-learning to adaptively adjust task weights and improve the positive effect of the auxiliary task. The method can learn refined expression representations latent in the facial topology. Experiments prove the effectiveness of the proposed method.

1. Introduction

Emotional states have been studied for a long history since the ancient Greek era [1], and it is widely believed that emotional states can be conveyed by facial expressions [5, 12, 19]. Therefore, sensing people's emotional states through facial expressions by computers arouses great interest among researchers. Many algorithms are based on the categorical model, though many other models are proposed to describe people's emotions [38]. Previous researches often focus on the model that classifies expressions into surprise, fear, disgust, happiness, sadness, and anger [11]. Recent studies have focused more attention on it that people would express more complex sentiments that are out of these six basic emotions [26, 31].

All facial expressions are composed of different facial muscles, i.e., action units (AUs) [10]. The facial action coding system (FACS) encodes emotions as combinations of basic muscle activation and defines a set of action units (AUs) to describe specific muscle activations [8, 9, 13], which is regarded as AU analyses on basic emotions. Du *et al*. [6, 7] constructed their model based on the basic model by combining two basic emotions and selecting meaningful ones. Emotions are eventually classified into two categories: basic emotions and compound emotions, where the latter contains 14 compound emotions. AU analyses are thus expanded from basic emotions to compound emotions [7]. However, compound one is only done in the lab since there is still no database with manually-annotated compound emotions and AUs. With the urge to understand emotions better, we aim to improve the accuracy of FER by performing AU analysis on compound facial expressions in the open environment.

In this paper, compound expressions and their relationships to AUs in the real-world social environment are studied. Since [2] concluded that most compound emotions are combined by two basic emotions, our study focuses on the mentioned 14 combinations instead of more complex emotions. First, we construct a Real-world Affective Faces database of Compound Expressions (RAF-CE)[1] by mapping multi-label blended expressions in RAF-ML [26] to compound emotions and trimming the database to 4,549 images to align images with that of RAF-AU [39]. AU analysis is subsequently studied on in-the-wild compound emotions. Mainly focusing on the variation of AU activations in the wild, we found that real-world AU activations

[1] http://www.whdeng.cn/RAF/model2.html

differed from posed ones in both AU patterns and activation frequencies. To the best of our knowledge, RAF-CE is the first real-world database containing both manually-annotated compound emotions and AU labels.

We then propose a new meta-learning multi-task learning (MML) with the inspiration of [22,23,28] for compound expression recognition in the wild. Focusing on FER, our net chooses AU detection as an auxiliary task. To better utilize task relationships, an alignment loss is introduced. With the alignment loss, AU and FE predictions constrain each other's distribution explicitly through AU activation statistics in AU analysis. Besides, to meet the task priority challenge of multi-tasking learning, meta-learning is adopted in our method. Keeping the dominant status of FER, the meta net automatically learns auxiliary task weights and constraint task weights. Experiments and ablation studies on RAF-CE show the effectiveness of the proposed method.

2. Related Work

2.1. Databases beyond facial expressions

Past research has proposed some databases limitedly focusing on lab-controlled basic images. Several databases broke through this limitation, but still not enough to be used for analyzing unconstrained compound emotions. We next mainly list these groundbreaking databases. RAF-DB [27] is a real-world database and proposes a compound emotion subset with 3,954 images among twelve compound emotions. AffectNet [33] collects more than one million unconstrained images with not only labels of eight basic emotions but also valence-arousal dimensions. Affwild2 [24] further considers expression labels in three dimensions, it contains 558 in-the-wild videos and most of the videos are manually annotated with expression labels, AU labels, and valence-arousal labels. However, it only focuses on 7 basic emotions. EmotioNet [14] is a large-scale database with one million images extracted from the Internet. Among these images, the AUs of 25,000 images were manually annotated while that of other images are automatically detected. A small subset of EmotioNet also provides 2,474 images with six basic expressions and ten compound expressions. However, these emotions are inferred from the images' AU labels.

2.2. Multi-task learning

Multi-task learning (MTL) is vastly similar to auxiliary learning and is a widely-known algorithm. The FER task is strongly related to the AU detection task. Emotions can be recognized through AU activations [13], and convolutional neural networks (CNNs) trained for FER tasks also can capture features that are highly related to AUs [21]. AU detection accordingly often acts as an auxiliary task to improve the FER performance [34].

Kollias and Zafeiriou [22, 23] formalized a new distribution matching loss to strengthen the relationship between the two tasks. It could accommodate AU detection with FER prediction via psychological lab-controlled AU analysis. Constrained by prior knowledge, AU prediction is tried to be consistent with FE prediction or vice versa. Nevertheless, this method ignores that the negative impact of prior constraints would increase as tasks are trained. Thus, we mitigate the excessive impact of alignment loss by adjusting the inputs of the loss.

A thorny problem is that imbalances between tasks may hinder proper training, since the network may assign the highest priority to one task. Therefore, MTL may excessively focus on one task by assigning gradients with larger magnitudes. Multiple MTL methods have been proposed to alleviate the problem [4, 15, 20, 25]. Our method uses meta-learning and adaptively weights the tasks' loss owing to tasks' sensitivity to weights [20], aiming to automatically adjust the importance of auxiliary tasks.

2.3. Meta auxiliary learning

Meta-learning is a general term describing the network's autonomically using prior experience to optimize itself. The 'learning-to-learn' [36] has been gaining attention recently as an approach that can solve some bottlenecks faced by machine learning and continuously advance deep learning research [17, 32]. Though not distilling experiences on other tasks [37], MTL benefits from combing with meta-learning [18, 29]. Our method is inspired by meta auxiliary learning (MAL) network [28]. The MAL network assigns adaptive weights to input samples from AU and FE datasets via meta-learning and works well for learning over different databases. However, our experiments show that the method it not so applicable to databases that concurrently have AU and FE labels.

3. Database

3.1. RAF-CE

Our dataset focuses on 14 compound emotions as well as 32 AUs. To build a dataset that could be used to analyze the AU activation of each compound emotion, RAF-ML [26] and RAF-AU [39] are chosen. RAF-ML is a facial expression dataset that provides nearly 5,000 in-the-wild images with multi-labels, where each class label stands for one basic emotion. Based on RAF-ML, Yan *et al*. [39] annotated AUs for these images to create RAF-AU. We first selected the intersection of the two databases, then mapped multi-label expressions to 14 compound expressions. Herein, we reviewed the distributions of each image's labels and retained the two labels with the highest probabilities. During the mapping process, images that are out of these compound emotions categories after the proposed process are

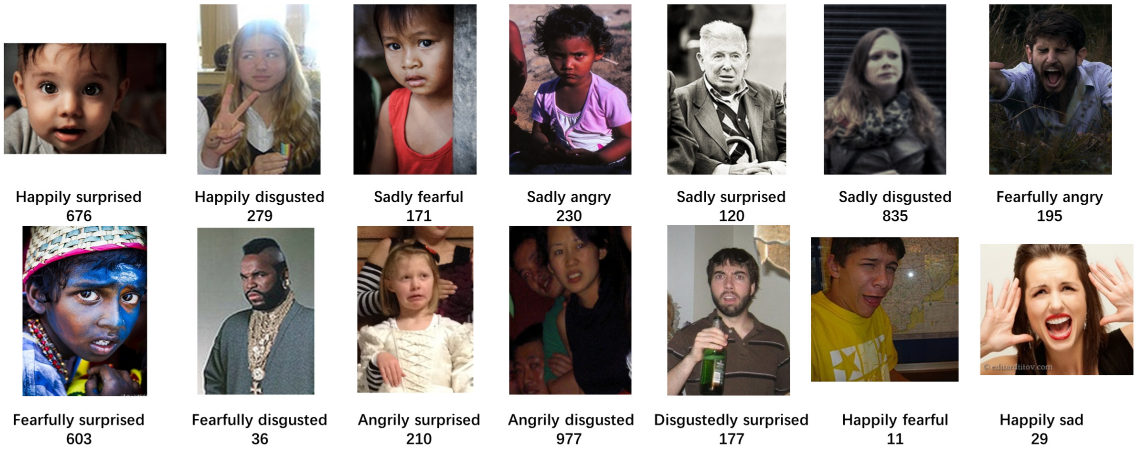


Figure 1. Examples of compound emotions in RAF-CE. Quantities of images are listed below pictures.

also eliminated, such as happily angry. 52 images are discarded, and 4549 images remain in total. Figure 1 demonstrates some pictures of these 14 compound emotions.

3.2. AU analysis

Herein, the AU activations and compound emotions in the wild are mainly focused. By using the AUs in RAF-AU to make the analysis, we count the frequencies of each AU presented over emotion, as shown in Tab. 1. It is necessary to mention that the three categories, fearfully disgusted, happily sad, and happily fearful, only have 36, 29 and 11 images in our database respectively. The AU analysis of these three emotions may vary as the number of images increases.

It is unsurprising that most AUs observed in each emotion are similar to the examination in [7]. However, differences are also relatively noticeable, reflected both in frequencies and AU patterns. For each emotion, Du *et al*. [6,7] classified activated AUs into prototypical AUs and variant AUs by the frequencies of subjects using each AU. AUs with frequencies equal to or greater than 70% are called prototypical AUs, and AUs with frequencies equal to or greater than 20% but less than 70% are called variant AUs. Besides, both prototypical AUs and variant AUs are considered as observed AUs.

Observation 1: Variation of observed AU patterns.

The most obvious change is the AU patterns of activation. People use different AUs when expressing their feelings in the wild. For example, AU 7 (lid tightener), which is prototypical in anger and is variant in all compound emotions that have anger as a subordinate class, is rarely activated in our database. Similarly, few people use AU 15 (lip corner depressor) to express their sadness, although the AU could be detected in the lab. Also, AU 20 (lip stretcher) is less activated when humans express feelings related to surprise and fear in the wild. Besides, AU 17 (chin raiser) and AU 24 (lip pressor) are also less activated in the open environment. Moreover, none of the expressions could have these three AUs as their observed AUs. Interestingly, AU 11(nasolabial deepener), as a variant AU in four lab-controlled compound emotions, is never detected in one face.

Conversely, people habitually activate other AUs like AU16 (lower lip depressor) and AU27 (mouth stretch). Both AUs are not listed as observed AUs in the lab experiment but are classified as variant AUs or prototypical AUs in RAF-CE. An interesting thing is that most of the changes in activation patterns happen in variant AUs, meaning most prototypical AUs can still indicate emotions. Figure 2 shows the exclusive observed AUs of two different environments.

In addition to the AUs whose activation states are highly correlated with the environment, there are also some AUs that differ greatly between pairs of expressions but do not show a high correlation with the environment. Take AU2 (outer brow raiser) as an example. In some expressions like sadly fearful, AU 2 tends to be observed in lab-controlled environments. Nevertheless, in other expressions like fearfully angry, it tends to be more activated in the wild. AU6 (cheek raiser), AU9 (nose wrinkler), and AU26 (jaw drop) have similar situations.

Contrary to our conjecture, spontaneous emotions do not involve more AU classes. Real-world emotions only have 13 observed AU classes in total while posed ones have 15 AU classes.

Observation 2: People habitually activated several AUs.

Another impressive thing is that people habitually use AU 12 (Lip Corner Puller) and AU 25 (Lips part) in their spontaneous emotions. Although thought as an indicator of

Emotion	Observed AUs 1	2	4	5	6	7	9	10	11	12	15	16	17	20	24	25	26	27
HSur.	40%	39%		32%						58%						**86%**	43%	29%
HSur.*	**95%**	**93%**		64%						**100%**						**100%**	67%	
HD			27%		21%		24%	38%		52%						50%		
HD*			32%		61%		59%	**98%**		**100%**						**100%**		
SF	40%		58%							40%		21%				**78%**	35%	25%
SF*	**86%**	46%	**94%**	24%	34%						30%			**70%**		**97%**		
SA			**71%**		20%		29%	29%		22%		26%				60%		26%
SA*			**97%**		26%	48%			20%		**83%**		50%					
SS	34%		37%													**71%**	47%	
SS*	**84%**	27%	**90%**		31%											**99%**	**90%**	
SD	20%		59%					30%					29%			26%		
SD*	49%		**97%**		61%		20%	**85%**	35%		54%		47%			43%		
FA		20%	47%	29%	27%		56%	**74%**		60%		63%				**98%**		**80%**
FA*			**99%**	40%		39%		30%	33%					**84%**		**98%**		
FS	50%	41%		64%						32%		26%				**84%**	33%	34%
FS*	**93%**	**80%**	47%	**74%**				35%	22%					**90%**		**99%**	51%	
FD	28%		56%	25%				53%		22%						**75%**	22%	
FD*	**77%**	64%	**75%**	50%	26%		28%	**92%**			33%			**88%**		**98%**		
AS	23%	23%	33%	38%				35%		23%		38%				**86%**	39%	36%
AS*			**99%**	35%		50%		34%								**100%**	**94%**	
AD			55%				33%	45%								44%		
AD*			**98%**			60%	57%	**93%**					**79%**		36%			
DS	30%	24%	38%	23%				35%								**75%**	39%	
DS*	**93%**	**90%**	45%	**73%**			37%	**91%**					66%		33%			
HF					27%			64%		**73%**						**73%**		55%
HF*	**90%**	**85%**		30%						**100%**						**100%**	**100%**	
HSad	28%		38%		24%			24%		52%						52%		
HSad*	**95%**	**93%**		64%						**100%**						**100%**	67%	

Table 1. Compound AU analysis. Lab-controlled AU analysis is also listed as a comparison. Emotion labels represent the following 14 expressions in order: happily surprised, happily disgusted, sadly fearful, sadly angry, sadly surprised, sadly disgusted, fearfully angry, fearfully surprised, fearfully disgusted, angrily surprised, angrily disgusted, disgustedly surprised, happily fearful, and happily sad. Emotion labels with stars indicate that these rows of data are lab-controlled and data in the other rows are from RAF-CE. AUs in bold are prototypical AUs. Underlined AUs mean the AUs are exclusively observed in unconstrained emotions (red) or in posed emotions (blue).

positive emotions, AU 12 is observed in 10 compound emotions. However, AU 12 can be observed in the lab only if the compound emotion has a happiness component: happily surprised, happily disgusted, happily fearful, and happily sad. Furthermore, AU 25 is observed frequently in the lab, with 11 times over 14 emotions. This proportion strikingly increases to 100% in the wild.

Observation 3: Reductions of AU activation frequencies.

Generally, the frequencies of AUs activated by people in the wild are less than that in the lab. With the drop in frequencies, many prototypical AUs are reclassified as variant AUs, especially for AU 1 (inner brow raiser) and AU 4 (brow lowerer). Similarly, some variant AUs like AU 2 and AU17 in the lab exist with a proportion less than 20% and thus cannot be classified as observed AUs. Nevertheless, there are also some newly-classified variant AUs, causing the rise of variant AU numbers. In summary, the mean number of prototypical AUs activated among 14 lab-controlled emotions is 3.30, while that of in-the-wild expressions is only 0.73. The average numbers of variant AUs are 1.35 and 2.12, respectively. As for all observed AU, 2.85 AUs are averagely activated in each spontaneous emotion compared with 4.65 AUs in each lab-controlled emotion. One exception is fearfully angry. 5.54 AUs are averagely activated to express fearfully angry, compared with 3.9 AUs in the lab.

4. Meta multi-task network

For a given image x, we first introduce the traditional MTL [3] network. For AU detection, binary cross-entropy loss is exploited, which is formulated as:

$$\begin{aligned} L^{AU} = -\sum_{j=1}^{32} & y_j^{AU} \log(p_j^{AU}) \\ & + (1 - y_j^{AU}) \log(1 - p_j^{AU}), \end{aligned} \tag{1}$$

where y_j^{AU} means ground truth AU annotation and p_j^{AU} means the prediction of jth AU.

For FER, the cross-entropy loss is chosen. It is formulated as:

$$L^{emo} = -\sum_{k}^{14} y_k^{emo} \log(p_k^{emo}), \tag{2}$$

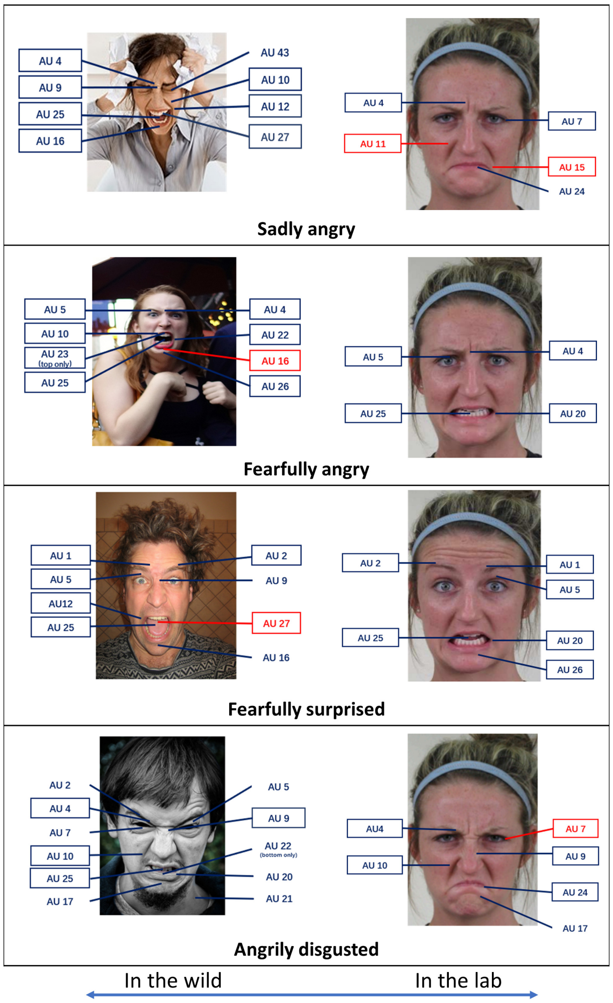


Figure 2. Comparisons of AU in different environments. Bounding boxes mark the observed AUs and red ones highlight the AUs that only exist in one database.

where y_k^{emo} and p_k^{emo} indicate kth ground truth and predicted score over 14 expressions, respectively. And for the traditional MTL, the network is trained to minimize the following multi-task loss:

$$L^{TMTL} = L^{emo} + \rho L^{AU}, \tag{3}$$

where ρ denotes the manually-set weight of AU loss.

4.1. Task explicit association

The distribution matching method aims to couple the two tasks explicitly. For FE prediction p^{emo} of an input x, its empirical AU distribution t^{AU} can be calculated from the statistics in Tab. 1:

$$t_j^{AU} = \sum_{k=1}^{14} p_k^{emo} p(y_j^{AU} | y_k^{emo}), \tag{4}$$

where $p(y_j^{AU} | y_k^{emo})$ is a constant that is derived from statistical results.

While [22, 23] used lab-controlled results, we choose in-the-wild ones. Once y_k^{emo} is fixed, $p(y_j^{AU} | y_k^{emo})$ would turn to statistical frequencies of corresponding AUs. For instance, y_9^{AU} denotes AU 9 and it can be observed in happily disgusted, sadly angry, fearfully angry, and angrily disgusted. Therefore, $t_9^{AU} = 0.24 \cdot p_{happily\_disgusted} + 0.29 \cdot p_{sadly\_angry} + 0.56 \cdot p_{fearfully\_angry} + 0.33 \cdot p_{angrily\_disgusted}$. Nevertheless, [22, 23] ignore the circumstance where p^{emo} will be extremely close to 1 for one emotion and very close to 0 for others as the FER is trained. t^{AU} then approximates the theoretical AU distribution of the emotion, which obstructing the learning of AU tasks. Therefore, the sigmoid function is adopted:

$$q_j^{AU} = \sum_{k=1}^{14} p_k^{emo} \cdot \mathrm{Sigmoid}(p(y_j^{AU} | y_k^{emo})). \tag{5}$$

In this case, when $p_k^{emo} = 1$, its corresponding q^{AU} will be larger and the rest will become smaller, but in the reasonable gap. The first distribution matching loss is subsequently adopted based on cross-entropy loss:

$$L^{DM1} = -\sum_{j=1}^{32} p_j^{AU} log(q_j^{AU}). \tag{6}$$

By minimizing the above loss, two distributions are led to match each other. It reflects the thought that, for example, when $p_9^{AU} = 1$, the emotion is more likely to be one of the four expressions mentioned above.

Another way is to calculate q_k^{emo} based on predicted AUs:

$$q_k^{emo} = \frac{\sum_{j=1}^{32} p_j^{AU} p(y_j^{AU} | y_k^{emo})}{\sum_{j=1}^{32} p(y_j^{AU} | y_k^{emo})}. \tag{7}$$

With q^{emo}, the second distribution matching loss is introduced :

$$L^{DM2} = -\sum_{k=1}^{14} p_k^{emo} \log(q_k^{emo}). \tag{8}$$

By combining two losses, an alignment loss is formulated:

$$L^{AM} = L^{DM1} + L^{DM2}. \tag{9}$$

In this stage, the alignment MTL network is trying to minimize the following loss:

$$L^{AMTL} = L^{emo} + \rho_1 L^{AU} + \rho_2 L^{AM}. \tag{10}$$

4.2. Meta multi-task learning

Our base net needs to perform three tasks simultaneously. The meta-learning network is utilized to automatically weigh each loss. Considering that FER should be the

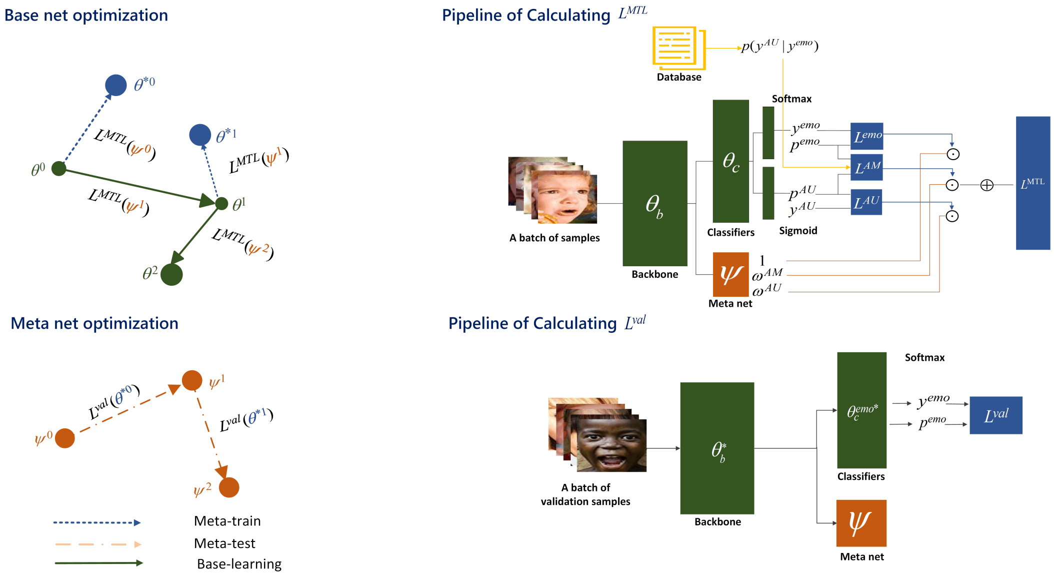


Figure 3. Pipeline of optimization. The left side is the pipeline of optimization, and the right side is the calculation diagram of losses. (i) Meta-train: The optimization direction of the base net θ^* is obtained by L^{MTL}. (ii) Meta-test: Meta net optimizes ψ according to L^{val} and θ^*. (iii) Base-learning: L^{MTL} is recalculated based on the updated ψ^*, then the base net is truly updated.

dominant task, we imposed a constraint that the weight of FER is one. The network learns to minimize the alignment multi-task loss:

$$L^{AMTL} = L^{emo} + \omega^{AU} L^{AU} + \omega^{AM} L^{AM}, \tag{11}$$

where ω^{AU} and ω^{AM} are adaptive weights learned by meta net.

As shown in Fig. 3, it contains two parts. One is the base net for predicting FEs and AUs of the images. The base net comprises a backbone net, a FE classifier, and an AU classifier. $\theta = \{\theta_b, \theta_c^{emo}, \theta_c^{AU}\}$ indicate parameters of the base net. We also denote $f(x)$ as outputs of the backbone net. The other part, denoted as $g(f(x);\psi)$, is the meta net with parameters ψ for predicting weights $\omega = \{w^{AU}, w^{AM}\}$. The training process of each iteration contains three parts: (i) meta-train, (ii) meta-test (iii) base-learning.

(i) Meta-train. As shown in Fig. 3, for a sample x_i in a mini-batch B from the training set, meta net predicts weights using features extracted by the backbone net. Combined with the outputs of the base net, the total loss will be:

$$L^{MTL} = \sum_{i=1}^{B} (L_i^{emo} + \omega_i^{AU} L_i^{AU} + \omega_i^{AM} L_i^{AM}). \tag{12}$$

The parameter will then be updated:

$$\theta^* = \theta - \alpha \nabla_\theta L^{MTL}, \tag{13}$$

where α indicates the learning rate of the base net.

(ii) Meta-test. This stage aims to optimize ψ based on the updated base net. On the validation set, we calculate the corresponding loss: $L^{val} = L^{emo}$. Then meta net will be updated:

$$\hat{\psi} = \psi - \beta \nabla_\psi L^{val} \left(\theta^{emo*}\right), \tag{14}$$

where β is learning rate of meta net, $\theta^{emo*} = \{\theta_b^*, \theta_c^{emo*}\}$ are updated parameters in base net. Herein, the network is to calculate the second derivative:

$$\nabla_\psi L^{val} = \nabla_\psi \left(\nabla_{\theta^{emo*}} L^{val}\right). \tag{15}$$

(iii) Base-learning. With newly-updated parameters of meta net, base net will be trained again by calculating $\hat{L}^{MTL}$.

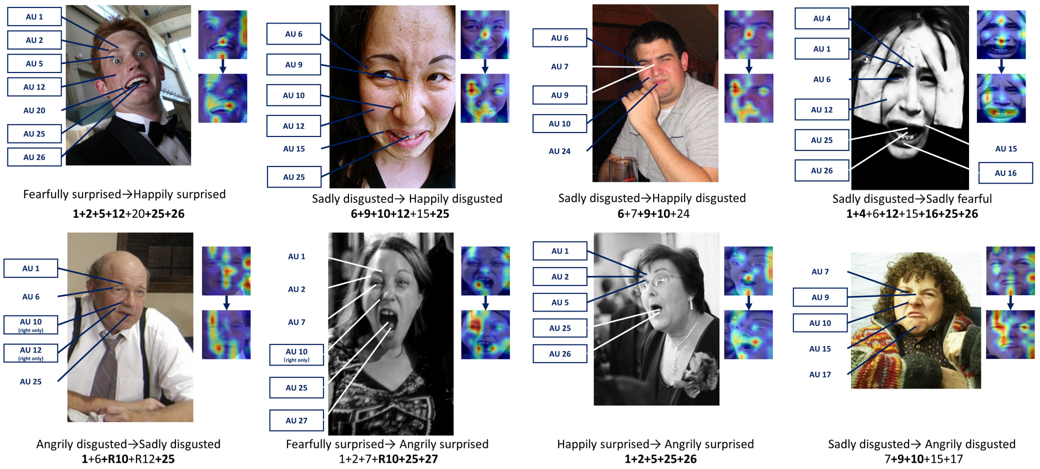
$$\begin{aligned} \hat{L}^{MTL} =& \sum_{i=1}^{B} L_i^{emo}(\theta^{emo}) + \sum_{i=1}^{B} \hat{w}_i^{AU} L_i^{AU}\left(\theta^{AU}\right) \\ &+ \sum_{i=1}^{B} \hat{w}_i^{AM} L_i^{AM}(\theta), \end{aligned} \tag{16}$$

where $\hat{\omega}_i^{AM}$ and $\hat{\omega}_i^{AU}$ means weights generated by newly-updated meta net.

Parameters in base net will be updated again:

$$\hat{\theta} = \theta - \alpha \nabla_\theta \hat{L}^{MTL}. \tag{17}$$

After finishing the training process, our net conducts testing by predicting only compound emotions.



Fearfully surprised→Happily surprised
1+2+5+12+20**+25+26**

Sadly disgusted→ Happily disgusted
6+9+10+12+15**+25**

Sadly disgusted→Happily disgusted
6+7**+9+10**+24

Sadly disgusted→Sadly fearful
1+4+6**+12**+15**+16+25+26**

Angrily disgusted→Sadly disgusted
1+6**+R10**+R12**+25**

Fearfully surprised→ Angrily surprised
1+2+7+**R10+25+27**

Happily surprised→ Angrily surprised
1+2+5+25+26

Sadly disgusted→ Angrily disgusted
7**+9+10**+15+17

Figure 4. Examples of pictures with corrected predictions by meta multi-task network, attached with different attention regions from traditional MTL and our method. AUs in boxes are observed AUs, which assist our methods to recognize emotions.

5. Experiments

5.1. Database

To our best knowledge, since RAF-CE is the first in-the-wild database with both compound emotion labels and AU manual annotations, our experiments were consequently mainly conducted on RAF-CE.

5.2. Training details

The training set was randomly selected by one-fourth to form the validation sets. A 34-layer ResNet [16] pre-trained on CASIA-WebFace [40] without the last fully connected layer was adopted as the backbone. It takes in an RGB image with 112×112 size and passes the output to classifiers and meta net. One fully connected layer for each classifier and one fully connected layer with a sigmoid activation function for meta net are adopted, where the parameters for the layer used in the meta net are set to zero. For RAF-CE, the mini-batch size is set to 16. Stochastic gradient descent (SGD) is chosen as an optimization method for base net and meta net. Learning rates of backbone, classifiers, and meta-net are 0.001, 0.01, and 0.01, respectively. Moreover, all our experiments are implemented using PyTorch on NVIDIA Tesla T4 GPU, and it costs approximately 15 hours to train the model. For single-task learning (STL) and traditional MTL, the pre-trained 34-layer ResNet is adopted. All network details are the same as that of the base net. Moreover, traditional MTL only calculates L^{emo} and binary cross-entropy loss L^{AU} with traversal AU weights.

5.3. Results

The overall accuracy was used as the main evaluation criterion. Besides, we also adopt the mean diagonal value of the confusion matrix over all emotions to evaluate the results.

Comparison with other methods. As shown in Tab. 2, it can be observed that our method gets better performance. Our method outperforms single-task learning and multi-task learning by 3.74% and 1.98%. Among 14 compound emotions, our method also achieves better results, particularly showing superior performance on categories with small sample sizes. This could be due to the fact that people express these emotions less frequently, resulting in relatively lower variability, and thus prior knowledge can have a greater impact.

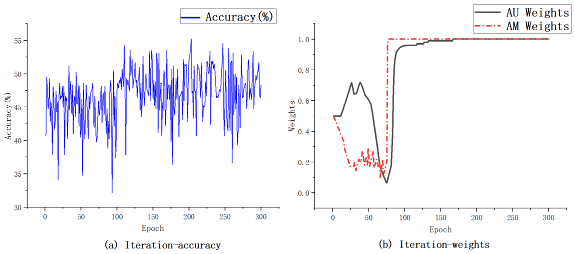
We select some samples, the recognition result of which are corrected by our method, and apply Grad-CAM [35] to visualize the corresponding attention maps of traditional multi-task learning and our method. As illustrated in Fig. 4, our net focuses more on the highly-related AUs to predict the correct class consequently.

Comparison with lab-controlled statistics. The relationships extracted by [6, 7] are also attempted in our methods. Some changes are made to get better results. Since categories of AUs vary more between different expressions, all $p(y_j^{AU}|y_i^{emo})$ are set to ones while moving the sigmoid function backward to balance alignment and AU detection: $q_j^{AU} = \text{Sigmoid}(\sum_{i=1}^{14} p_i^{emo} \cdot p(y_j^{AU}|y_i^{emo}))$. The results are shown in Tab. 3, where the method with our statics outperforms theirs by 0.22%.

Ablation study. We also analyze our method using ablation studies by using meta-learning or aligning distributions individually. Two meta nets for meta-learning are trained to choose the better one. One only predicted ω^{AU}, with $L^{total} = L^{emo} + \omega^{AU} L^{AU}$. The other predicted both ω^{AU} and ω^{emo}, following the loss $L^{total} = \omega^{emo} L^{emo} + \omega^{AU} L^{AU}$. For alignment loss method, the net with $L^{total} = L^{emo} + \rho_1 L^{AU} + \rho_2 L^{AM}$ is trained.

	happily surprised	happily disgusted	sadly fearful	sadly angry	sadly surprised	sadly disgusted	fearfully angry	fearfully surprised	fearfully disgusted	angrily surprised	angrily disgusted	disgustedly surprised	happily fearful	happily sad	Acc.	Diag.
STL	57.14	**18.52**	11.11	9.09	12.00	**73.03**	39.53	**78.76**	0	**29.55**	64.95	0	0	0	51.49	28.12
MTL	62.7	**18.52**	8.33	11.36	**24.00**	71.35	**41.86**	74.34	0	27.27	71.65	2.70	0	0	53.25	29.58
MML(Ours)	**77.78**	7.41	**27.78**	**13.64**	20.00	62.92	**41.86**	61.95	0	25.00	**84.54**	**8.11**	0	**12.5**	**55.23**	**31.68**

Table 2. Performance comparison on RAF-CE. Accuracies on each emotion are listed. 'Acc.' is the overall accuracy; 'Diag.' is the mean diagonal value of the confusion matrix.



(a) Iteration-accuracy

(b) Iteration-weights

Figure 5. Curves: (a) Iteration-accuracy (b) Iteration-weights

Statistics	Method	Accuracy	Diag.
	MAL	52.37	30.73
Lab [6, 7]	AM	54.12	30.15
	MML(Ours)	55.01	30.85
In the wild	AM	53.36	28.02
	MML(Ours)	**55.23**	**31.68**

Table 3. Ablation study on RAF-ML. 'MAL' means meta auxiliary network is separately used. 'AM' means alignment loss method is separately used. 'Diag.' means the mean diagonal value of the confusion matrix.

As shown in Tab. 3, our method outperforms meta auxiliary learning and alignment net by 2.86% and 1.87%. Furthermore, MAL method underperforms MTL, when used individually. The poor results of MAL reflect that paying too much attention to AU detection without constraints will cause negative transfer [30]. Also, statistics from labs perform better when we use alignment loss individually. The disappointing result is related to the low proportion of AU activation that we summarized earlier. For in-the-wild statistics, premature introduction of the alignment loss can inhibit AU detection.

Weights analysis. Changes in weights by calculating average weights in each epoch are recorded. The statistics are visualized in Fig. 5. ω^{AM} first decreased and next increased rapidly to one, and ω^{AU} firstly showed an upward trend with some fluctuations, then decreased sharply, followed by an increment. One can observe that the changes between the two weights are roughly opposite during the early stages. The phenomenon shows that the net first focused on the FER task. After the convergence of the FER task, AU detection and alignment turned into dominant tasks. As the FER accuracy rose, the alignment loss had a positive effect and improved the accuracy of FER. And before FER task is fully trained, alignment of AU and expression distribution will hinder task coordination.

6. Conclusion

This work simplified the real-world multi-label facial expression database RAF-ML to a compound expression database RAF-CE. AU analysis was conducted with manually annotated AUs extracted from RAF-AU. Moreover, AU analysis has shown the diversity and differences in real-world muscle activation from that in the lab. Based on the analysis, we presented a meta-based multi-task network to get better collaboration between AU detection and FER tasks. Experiments and visualized results confirmed the effectiveness of the method.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grant No.62276030 and 62236003, China National Postdoctoral Program for Innovative Talents (Grant No. BX2021047) and China Postdoctoral Science Foundation (Grant No. 2022M710466).

References

[1] Joan C Borod et al. *The neuropsychology of emotion*. Oxford University Press, 2000. 1

[2] De'Aira Bryant, Siqi Deng, Nashlie Sephus, Wei Xia, and Pietro Perona. Multi-dimensional, nuanced and subjective-measuring the perception of facial expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20932–20941, 2022. 1

[3] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 4

[4] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018. 2

[5] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. 1

[6] Shichuan Du and Aleix M Martinez. Compound facial expressions of emotion: from basic research to clinical applications. *Dialogues in clinical neuroscience*, 2022. 1, 3, 7, 8

[7] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014. 1, 3, 7, 8

[8] Paul Ekman. Pictures of facial affect. *Consulting Psychologists Press*, 1976. 1

[9] Paul Ekman. Methods for measuring facial action. *Handbook of methods in nonverbal behavior research*, pages 45–90, 1982. 1

[10] Paul Ekman. Are there basic emotions? 1992. 1

[11] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. 1

[12] Paul Ekman. *Darwin and facial expression: A century of research in review*. Ishk, 2006. 1

[13] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1, 2

[14] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016. 2

[15] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287, 2018. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[17] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021. 2

[18] Dasol Hwang, Jinyoung Park, Sunyoung Kwon, Kyung-Min Kim, Jung-Woo Ha, and Hyunwoo J Kim. Self-supervised auxiliary learning with meta-paths for heterogeneous graphs. *Advances in Neural Information Processing Systems*, 33:10294–10305, 2020. 2

[19] William James. Discussion: The physical basis of emotion. *Psychological review*, 1(5):516, 1894. 1

[20] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 2

[21] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE international conference on computer vision workshops*, pages 19–27, 2015. 2

[22] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 2, 5

[23] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 2, 5

[24] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023. 2

[25] Changsheng Li, Junchi Yan, Fan Wei, Weishan Dong, Qingshan Liu, and Hongyuan Zha. Self-paced multi-task learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[26] Shan Li and Weihong Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowd-sourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127(6):884–906, 2019. 1, 2

[27] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 2

[28] Yong Li and Shiguang Shan. Meta auxiliary learning for facial action unit detection. *IEEE Transactions on Affective Computing*, 2021. 2

[29] Xingyu Lin, Harjatin Baweja, George Kantor, and David Held. Adaptive auxiliary task weighting for reinforcement learning. *Advances in neural information processing systems*, 32, 2019. 2

[30] Shengchao Liu, Yingyu Liang, and Anthony Gitter. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9977–9978, 2019. 8

[31] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A

large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2208.00847*, 2022. 1

[32] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018. 2

[33] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 2

[34] Gerard Pons and David Masip. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*, 2018. 2

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7

[36] Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pages 3–17. Springer, 1998. 2

[37] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018. 2

[38] Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. A review of emotion sensing: categorization models and algorithms. *Multimedia Tools and Applications*, 79(47):35553–35582, 2020. 1

[39] Wen-Jing Yan, Shan Li, Chengtao Que, Jiquan Pei, and Weihong Deng. Raf-au database: in-the-wild facial expressions with subjective emotion judgement and objective au annotations. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 2

[40] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 7