RE-Net: A Relation Embedded Deep Model for AU Occurrence and Intensity Estimation

Huiyuan Yang[0000−0003−0517−5187] and Lijun Yin[0000−0002−0343−7190]

Department of Computer Science
Binghamton University-State University of New York, Binghamton, NY, USA
hyang51@binghamton.edu, lijun@cs.binghamton.edu

Abstract. Facial action units (AUs) recognition is a multi-label classification problem, where regular spatial and temporal patterns exist in AU labels due to facial anatomy and human's behavior habits. Exploiting AU correlation is beneficial for obtaining robust AU detector or reducing the dependency of a large amount of AU-labeled samples. Several related works have been done to apply AU correlation to model's objective function or the extracted features. However, this may not be optimal as all the AUs still share the same backbone network, requiring to update the model as a whole. In this work, we present a novel AU **R**elation **E**mbedded deep model (**RE-Net**) for AU detection that applies the AU correlation to the model's parameter space. Specifically, we format the multi-label AU detection problem as a domain adaptation task and propose a model that contains both shared and AU specific parameters, where the shared parameters are used by all the AUs, and the AU specific parameters are owned by individual AU. The AU relationship based regularization is applied to the AU specific parameters. Extensive experiments on three public benchmarks demonstrate that our method outperforms the previous work and achieves state-of-the-art performance on both AU detection task and AU intensity estimation task.

1 Introduction

Automatic facial action units (AUs) recognition has attracted increasing attention in recent years due to its wide-ranging applications in affective computing, social signal processing, and behavioral science. Based on the Facial Action Coding System (FACS) [1], action units which refer to the contraction or relaxation of one or more facial muscles, have been used to infer facial behaviors for emotion analysis.

Automatic AU recognition is a challenging task due to many factors, such as image conditions, size of database, and individual differences. Although large-scale training data can facilitate the learning process for AU classification, data collection and AU annotation are extremely labor-intensive, thus being a time-consuming and error prone process. Fortunately, behavior research shows that there exist regular spatial and temporal patterns in AU labels due to facial anatomy and human's behavior habits. For example, persons can not *pull lip*

Fig. 1. Coupling effect of multiple AUs caused by a variety of facial expressions.

corner (AU12) and *depress lip corner* (AU15) at the same time due to the constraint of facial anatomy; *inner brow raiser* (AU1) and *outer brow raiser* (AU2) are both related to the muscle group *Frontalis*, most people cannot make a facial movement of AU1 without AU2, and vice versa. Fig.1 shows the coupling effect of multiple AUs caused by a variety of facial expressions. Such regular spatial and temporal patterns embedded in AU labels could be used as a constraint for AU detection.

Inspired by the above observations, there has been extensive research by exploiting AU relations to facilitate the learning process of AU classifiers. For example, the existing works reported in [2][3][4][5] have proposed to apply graphical model to capture the dependencies among AUs through its structure and conditional probabilities. Alternatively, recent works by Benitez-Quiroz et al. [6], Zhang et al. [7] and Peng et al. [8][9] proposed to introduce the dependencies among AUs into the objective function, and Zhao et al. [10][11] exploited the relationship among AUs and active image patches. To further utilize the AU relationships, Corneanu et al. [12], Li et al. [13], and Shao et al. [14] proposed to extract robust features by adding the AU relationship based graph neural networks to the extracted features.

However, the existing AU relationship modeling (at the *objective function level*, *feature level*, and *image patch level*) may not be optimal. First, most of the works that model the AU relationship as a prior rule into the classification predictions are usually not end-to-end trainable; Second, all the AUs still share the same backbone network, which is updated as a whole, thus it may not be the optimal way to utilize the AU relationship. As a matter of fact, the AU relationship can not be effectively used to update the model's parameters because the AU relationship is only involved into the calculation of the loss function. Moreover, in order to apply an AU-relation based graph to the extracted features, existing methods have applied a cropping operation to crop the AU-related region for

individual AU feature extraction, which may lead to the information loss due to the neglect of AU correlation. Importantly, all shared parameters of the model may not work equally well for different AUs.

Taking into account the shortcomings, we propose a new end-to-end trainable AU relation embedded deep model (**RE-Net**) that integrates the AU relationship into the model's parameter space. Specifically, instead of sharing the same backbone, different AUs contain both shared and AU-specific model parameters, in which the shared parameters are shared by all the AUs, and the AU-specific parameters are AU dependent. An AU relationship graph is constructed from the AU labels in the training dataset with AU as vertex and relationship as edge, which then used as a regularization of the AU-specific parameters. The benefits of using both shared and AU-specific model parameters with AU relationship are three-fold. First, by splitting the backbone parameters into both shared and AU-specific parts, the deep model is able to get updated both globally (through updating shared parameters for all the AUs) and locally (through updating AU-specific parameters for individual AU). Second, unlike existing methods that may lose information by cropping the facial region for individual feature extraction, our proposed method extracts different features from the input image for individual AU detection (as shown in Fig.3). Third, optimizing the AU-specific parameters by taking the AU relationship into account, our method is beneficial for recognition of AUs with less occurrence rate, potentially being capable of recognition of new AUs as well (more details in Section 4.6).

The contributions of this paper can be summarised as below:

- Built upon the adaptive batch normalization method, we format the multi-label AU detection problem as a domain adaptation problem with AU relation embedded, and propose a framework which contains both shared and AU-specific parameters.
- We conduct extensive experiments on the widely used datasets for both AU recognition and AU intensity estimation, and demonstrate the superiority of the proposed method over the state-of-the-art methods.
- Ablation study shows our model is extendable to recognize new AUs and robust to the scenario of data imbalance.

2 Related works

AU recognition is a multi-label classification problem, where multiple AUs may be present simultaneously, on the other hand, some AUs just can not happen at the same time. Exploiting AU relations has the potential to facilitate the learning process of AU classifiers.

Generative models are used to model the joint distribution. Li et al.[2] proposed to learn a dynamic Bayesian networks to model the relationships among AUs. Wang et al.[3] used restricted Boltzman Machine to capture both local pair-wise AU dependencies and global relationships among AUs. Tong et al.[4]

proposed Bayesian Networks to model the domain knowledge (AU dependencies) through its structure and conditional probabilities, and experimental results demonstrated that the domain knowledge can be used to improve parameter learning accuracy, and also reduced the dependency on the labeled data. Similar idea was also used in [5], which presents a learning algorithm to learn parameters in Bayesian networks under the circumstances that the training data is incomplete or sparse or when multiple hidden nodes exist.

On the other hand, discriminative approaches introduce the dependencies among AUs into the objective function; Zhao et al.[10] proposed a joint-patch and multi-label (JPML) method to exploit dependencies among AUs and facial features, which used the group sparsity and positive and negative AU correlations as the constraints to learn multiple AU classifiers. Zhao et al. [11] proposed a unified Deep Region and Multi-label Learning (DRML) network that simultaneously addresses both the strong statistical evidence of AU correlations and the sparsity of active AUs on facial regions. Peng and Wang [8] utilized the probabilistic dependencies between expressions and AUs as well as dependencies among AUs to train a model from partially AU-labeled and fully expression labeled facial images. Peng and Wang [9] used the dual learning method to model the dependencies between AUs and expressions for AU detection. By leveraging prior expression-independent and expression-dependent probabilities on AUs, Zhang et al. [7] proposed a knowledge-driven method for jointly learning multiple AU classifiers without AU annotations.

Instead of applying AUs dependencies into the objective function, some works also exploit using AU correlation as constraint for feature representation learning. Corneanu et al.[12] proposed a deep structured inference network (DSIN) to deal with patch and multi-label learning for AU recognition, which first extract local and global representations, and then capture AU relations by passing information between predictions using a graphical models. However, the relationship inference is still limited to the label level. Li et al. [13] proposed a AU semantic relationship embedded representation learning framework, which incorporate AU relationships as an extra guidance for the representation learning. A Gated Graph Neural Network (GCNN) is constructed using a knowledge-graph from AU correlation as its structure, and features extracted from facial regions as its nodes. As a result, the learned feature involves both the appearance and the AU relationship. A similar idea is also used in [14] that captures spatial relationships among AUs as well as temporal relations from dynamic AUs.

However, applying AU relationship to the objective function or the extracted features may not be optimal, so we propose to exploit the AU relationship in the model's parameter space.

3 Proposed method

3.1 Problem Formulation

The primary objective of our methodology is to learn a model with both shared and AU-specific parameters. To this end, our model seeks to learn the AU-specific

parameters in order to satisfy the AU correlation, and the shared parameters for AU detection. Formally, Let us consider $\mathcal{C}$ AUs to recognize. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \subset \mathcal{C}$ represents the set of vertices corresponding to AUs and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the set of edges, i.e., relations between AUs. In addition, we define an edge weight $\mathcal{W} : \mathcal{E} \to \mathbb{R}$ that measures the relation between two AUs. The network parameters are represented as $\theta = \{\theta_s, \theta_c\}_{c=1}^{\mathcal{C}}$,where θ_s represents the shared parameters for all AUs, while $\theta_c = (\gamma_c, \beta_c)$ represents the AU specific parameters owned by individual AU. Our goal is to learn a model with parameters of $\{\theta_s, \theta_c\}_{c=1}^{\mathcal{C}}$ by minimizing the supervised loss subject to the graph $\mathcal{G}$ among AU specific parameters $\{\theta_c\}_{c=1}^{\mathcal{C}}$.

3.2 Preliminary: Batch Normalization

A standard Batch Normalization (BN)[15] layer normalizes its input according to:

$$BN(x) = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{1}$$

where μ, σ^2 are the estimated mean and variance of x; γ, β are learnable scale and bias parameters respectively, and ϵ is a small constant used to avoid numerical instabilitie-s. For simplicity, the channel dimension and spatial location have been omitted.

Recent works[16][17][18][19] have shown the effectiveness of extending batch-normalization to address domain adaptation tasks. In particular, these works rewrite each BN to take into account domain-specific statistics. For example, given an AU $c \in \mathcal{C}$, a $\widehat{BN}$ layer differs from the standard BN by including a specific AU information:

$$\widehat{BN}(x, c) = \gamma \cdot \frac{x - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta \tag{2}$$

where $\{\mu_c, \sigma_c^2\}$ are the mean and variance statistics estimated from x conditioned on AU c. In other words, for each input x, the normalization is conditioned on which AU we aim to recognize. Since we do not want to share the scale and bias parameters across different AUs, so we include them within the set of private parameters, and rewrite the $\widehat{BN}(x, c)$ as below:

$$\widehat{BN}(x, c) = \gamma_c \cdot \frac{x - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c \tag{3}$$

here, $\{\gamma_c, \beta_c\}$ are the learnable AU specific parameters.

3.3 AU Relationship Graph

AU relationship graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents the correlations between each pair of AUs. Each node in graph represents a corresponding AU. Given a dataset with $\mathcal{C}$ AUs, the constructed graph is formed by $|\mathcal{C}|$ nodes.

Fig.2 shows the relation matrix studied on the datasets. Following previous work [14] [13] , we compute the Pearson correlation coefficient (PCC) between each pair of the j-th and i-th AUs in the dataset, denoted as $\omega_{i,j}$. Unlike [14] [13] that convert the AU correlations to positive or negative relationship based on two thresholds, ignoring the correlations between the two thresholds and also the strength of correlation, we use the original PCC as $\omega_{i,j}$, so both the positive/negative relationship and strength are considered.

3.4 AU Recognition with Graph Constraint

When training the model, $\widehat{BN}(x, c)$ allows to optimize the AU-specific scale γ_c and bia β_c parameters, however, it does not take into account the presence of the relationship between the AUs, as imposed by the AU correlation matrix. As used in [16], one possible way to include the AU correlation matrix within the optimization procedure is to modify Eq(3) as follows:

$$\widehat{BN}(x, c, \mathcal{G}) = \gamma_c^{\mathcal{G}} \cdot \frac{x - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c^{\mathcal{G}} \tag{4}$$

where, $\gamma_c^{\mathcal{G}}$ and $\beta_c^{\mathcal{G}}$ are calculated as below:

$$\gamma_c^{\mathcal{G}} = \frac{\sum_{k \in \mathcal{C}} \omega_{c,k} \cdot \gamma_k}{\sum_{k \in \mathcal{C}} \omega_{c,k}}; \qquad \beta_c^{\mathcal{G}} = \frac{\sum_{k \in \mathcal{C}} \omega_{c,k} \cdot \beta_k}{\sum_{k \in \mathcal{C}} \omega_{c,k}}; \tag{5}$$

$\omega_{c,k}$ is set as 1 if $c = k$, otherwise, $\omega_{c,k}$ represents the calculated PCC from training dataset. By doing this, the calculation of any AU-specific scale and bias parameters are influenced by other AUs with graph edge as the weight.

A cross-entropy loss function is used for AU recognitionn:

$$\mathcal{L}_\theta = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c \in \mathcal{C}} -\Big[y_{i,c}^T \times log(\bar{y}_{i,c}) + (1 - y_{i,c})^T \times log(1 - \bar{y}_{i,c}) \Big] \tag{6}$$

During training, we have two different strategies we can use:

- for each batch, we run the model $|\mathcal{C}|$ times to calculate the loss for each AU, and then update the model's parameters by back-propagating the sum of all the losses;
- for each batch, randomly select a single AU for optimization;

the first training method optimize the shared and all the AU specific parameters together, which may be beneficial for stable training, but the training procedure will be memory-intensive and time consuming. On the other hand, the second training strategy will not add extra burden by optimizing randomly selected single AU for each input batch, so the model can be trained as fast as the baseline model. Through experiments, we find that there is no big difference in performance as using two training strategies, so the second training method is of course preferred.

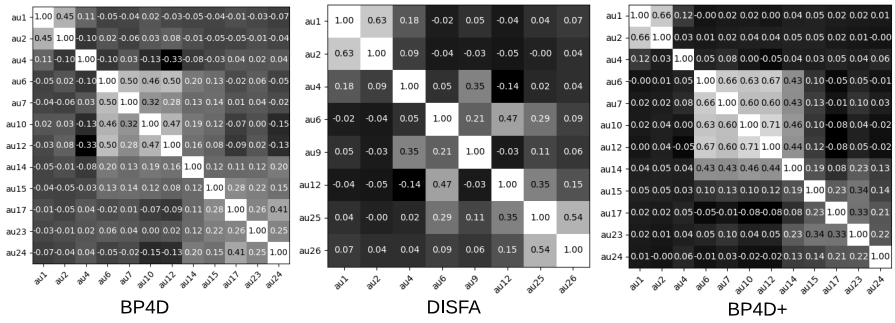

Fig. 2. The relation matrix calculated by PCC on three datasets. $(+, -)$ represents the corresponding positive and negative correlations between AU pairs; the absolute value means the strength of correlations. Zoom in for more details.

4 Experiments

In this section, the proposed method is first evaluated on three benchmark datasets: BP4D [20], DISFA [21] and BP4D+ [22] for AU recognition task, then applied to AU intensity estimation task in both BP4D and DISFA datasets.

4.1 Data

BP4D [20] is a widely used dataset for evaluating AU detection performance. The dataset contains 328 2D and 3D videos collected from 41 subjects (23 females and 18 males) under eight different tasks. As mentioned in the dataset, the most expressive 500 frames (around 20 seconds) are manually selected and labeled for AU occurrence from each one-minute long sequence, resulting in a dataset of around 140,000 AU-coded frames. For a fair comparison with the state-of-the-art methods, a three-fold subject-exclusive cross validation is performed on 12 AUs.

DISFA [21] is another benchmark dataset for AU detection, which contains videos from left view and right view of 27 subjects (12 females, 15 males). 12 AUs are labeled with AU intensity from 0 to 5, resulting in around 130,000 AU-coded images. Following the experimental setting in [23], 8 of 12 AUs with intensity greater than 0 are used from the left camera. F1-score is reported based on subject-exclusive 3-fold cross-validation.

BP4D+[22] is a multimodal spontaneous emotion dataset, where high-resolution 3D dynamic model, high-resolution 2D video,thermal (infrared) image and physiological data were acquired from 140 subjects. There are 58 males and 82 females, with ages ranging from 18 to 66 years old. Each subject experienced 10 tasks corresponding to 10 different emotion categories, and the most facially-expressive 20 seconds from four tasks were AU-coded from all 140 subjects, resulting in a database contains around 192,000 AU-coded frames. Following a similar setting in BP4D dataset, 12 AUs are selected and performance of 3-fold cross-validation is reported.

4.2 Evaluation Metrics

For the AU recognition task, we use the F1-score for comparison study with the state of the arts. F1-score is defined as the harmonic mean of the precision and recall. As the distribution of AU labels are unbalanced, F1-score is a preferable metric for performance evaluation.

For the AU intensity estimation task, we use the Intra-class Correlation ICC(3,1)[24], which is commonly used in behavioral sciences to measure agreement between annotators (in our case, the AU intensity levels between prediction and ground-truth). We also report the Mean Absolute Error (MAE), the absolute differences between target and prediction, commonly used for ordinal prediction tasks.

4.3 Implementation details

All the face images are aligned and cropped to the size of 256x256 using affine transformation based on the provided facial landmarks, randomly cropped to 224x224 for training, and center-cropping for testing. Random horizontal flip is also applied during training.

To analyze the impact of our proposed method, we use the ResNet-18[25] architecture as baseline. In particular, the default batch normalization layer is replaced with $\widehat{BN}(x, c, \mathcal{G})$ as described in Eq(4). To reduce the training complexity, a single AU is randomly selected to optimize for each training batch images, and use all of the AUs for validation and testing. We use an Adam optimizer with learning rate of 0.0001 and mini-batch size 100 with early stopping. We implement our method with the Pytorch[26] framework and perform training and testing on the NVIDIA GeForce 2080Ti GPU.

4.4 AU detection results

We compare our method to alternative methods, including Liner SVM (LSVM) [27], Joint Patch and Multi-label (JPML)[10], Deep Region and Multi-label (DRML) [11], Enhancing and Cropping Network (EAC-net)[23], Deep Structure Inference Network (DSIN) [12], Joint AU Detection and Face Alignment (JAA) [28],Optical Flow network (OF-Net) [29], Local relationship learning with Person-specific shape regularization (LP-Net) [30], Semantic Relationships Embedded Representation Learning (SRERL) [13] and ResNet18.

Table.1 shows the results of different methods on the BP4D database. It can be seen that our method outperforms all of the SOTA methods. The ResNet18 baseline achieves 59.6% F1-score, and a similar baseline performance is also reported in [31] [32]. Compared with the patch or region-based methods: JPML and DRML, our method achieves 19.6% and 17.2% higher performance on BP4D database. Compared with JAA and LP-Net, which used Facial landmarks as a joint task or regularization for AU detection, our method still shows 5.5% and 4.5% improvement in terms of F1-score on the BP4D database. It worth to note that both our method and LP-Net use ResNet as the stem network.

Table 1. F1 scores in terms of 12 AUs are reported for the proposed method and the state-of-the-art methods on the BP4D database. Bold numbers indicate the best performance; bracketed numbers indicate the second best.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	***Avg***
LSVM [27]	23.2	22.8	23.1	27.2	47.1	77.2	63.7	64.3	18.4	33.0	19.4	20.7	35.3
JPML[10]	32.6	25.6	37.4	42.3.	50.5	72.2	74.1	65.7	38.1	40.0	30.4	42.3	45.9
DRML[11]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
EAC-net[23]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
DSIN [12]	[51.7]	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	[62.9]	38.8	41.6	58.9
JAA [28]	47.2	44.0	54.9	[77.5]	74.6.	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
OF-Net [29]	50.8	[45.3]	[56.6]	75.9	75.9	80.9	[88.4]	63.4	41.6	60.6	39.1	37.8	59.7
LP-Net [30]	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	[45.3]	60.5	**48.1**	**54.2**	61.0
SRERL [13]	46.9	[45.3]	55.6	77.1	**78.4**	83.5	87.6	63.9	**52.2**	**63.9**	[47.1]	[53.3]	[62.9]
ResNet18	44.5	45.1	51.1	**81.2**	76.8	[87.6]	86.8	[67.9]	44.2	57.5	42.0	30.8	59.6
Ours	**57.7**	**59.0**	**66.9**	76.3	[77.0]	**88.9**	**89.8**	**70.9**	42.0	62.8	44.8	49.3	**65.5**

DSIN and SRERL are the closely related methods. Both DSIN and our method are able to predict label for individual AU, while SRERL and our method are similar in concept to learn robust feature for individual AU. The main difference lies in the facts: first, the CNN layers are still shared by all the AUs in both DSIN and SRERL, and the AU correlation is applied to the objective function or the extracted feature; second, DSIN needs incremental training and SRERL uses facial landmarks to crop the AU region for individual feature extraction, which may lead to information loss. Our end-to-end trainable method contains both shared and AU-specific parameters, so features can be extracted for individual AU by AU-relation guided computation of AU-specific parameters. The AU relationship is directly applied to the model's parameter space, and the 6.6% and 2.6% higher F1-scores demonstrate the effectiveness of applying AU relationship into the model's parameter space.

Experimental results on the DISFA database are reported in Table.2. As compared to ResNet18, our method shows 4.6% improvement. Note that, both JAA and LP-Net use facial landmarks as either a joint task or regularization, and SRERL uses AU intensity equal or greater than 2 as positive example, while our method and other methods use AU intensity greater than 0 as positive example, and our method still shows comparable result.

Our method is also evaluated on the BP4D+ database, which contains more AU-labeled frames from more subjects, the results are shown in Table.3. Our method achieves 4.0% improvement in F1-score when compared to ResNet18.

4.5 AU Intensity Estimation

Action Units recognition aims to detect the occurrence or absence of AUs; while, AU intensity is used to describe the extent of muscle movement, which presents detailed information of facial behaviours. AU intensity is quantified into six-point ordinal scales in FACS. Compared to AU detection, AU intensity estimation

Table 2. F1 scores in terms of 8 AUs are reported for the proposed method and the state-of-the-art methods on DISFA dataset. Bold numbers indicate the best performance; bracketed numbers indicate the second best. [* *means the method used AU intensity greater or equal to 2 as positive example.*]

Method	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	***Avg***
LSVM [27]	10.8	10.0	21.8	15.7	11.5	70.4	12.0	22.1	21.8
DRML [11]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-net [23]	41.5	26.4	66.4	**50.7**	**80.5**	**89.3**	88.9	15.6	48.5
DSIN [12]	42.4	39.0	[68.4]	28.6	46.8	70.8	90.4	42.2	53.6
JAA [28] *	[43.7]	[46.2]	56.0	41.4	44.7	69.6	88.3	58.4	[56.0]*
OF-Net [29]	30.9	34.7	63.9	44.5	31.9	[78.3]	84.7	[60.5]	53.7
LP-Net [30]*	29.9	24.7	**72.7**	46.8	49.6	72.9	[93.8]	**65.0**	**56.9***
SRERL [13] *	**45.7**	**47.8**	59.6	47.1	45.6	73.5	84.3	43.6	55.9*
ResNet18	31.3	33.7	48.7	45.5	33.3	68.6	**94.3**	48.1	50.4
Ours	38.8	31.1	57.2	[50.1]	[50.2]	75.5	86.6	50.6	55.0

Table 3. F1 scores in terms of 12 AUs are reported on the BP4D+ dataset.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	***Avg***
ResNet18	34.6	34.6	33.1	84.9	87.0	90.0	88.9	80.4	53.3	38.7	54.7	13.4	57.8
Ours	37.6	33.7	37.1	85.8	89.2	90.7	89.3	80.6	63.0	46.0	55.2	33.5	61.8

could provide more detailed information for facial behaviour analysis, but it is also a more challenging task, as the subtle difference among neighbor intensities.

Fortunately, a similar relationship also exists among AU intensity. We slightly modify the model for the AU intensity estimation task in both BP4D and DISFA datasets. First, AU intensity relationship is constructed from the AU intensity labels in the training dataset; second, the output of the model is set as six, the number of intensity levels. The results are shown in Table.4 and Table.5. Our method achieves the best average performance on both databases under two evaluation metrics, except ICC(3,1) on the BP4D database, where our method shows comparable result over the state-of-the-art methods.

4.6 Ablation study

Effectiveness of RE-Net: First, we provide evidences to support our claim that applying AU relation to the model's parameters space is more effective than applying to image patches or deep features. JPML[10] and DRML[11] model the AU relation with active image patches, while DSIN[12] and SRERL[13] applied the AU relation to the extracted deep features. As shown in Table.7, our method applies the AU relation to the parameter space of the model, which achieves

Table 4. ICC(3,1) and MAE scores in terms of 5 AUs are reported for the proposed method and the state-of-the-art methods on BP4D dataset. Bold numbers indicate the best performance; bracketed numbers indicate the second best.

	Method	AU6	AU10	AU12	AU14	AU17	Avg.
ICC	VGP-AE[33]	0.75	0.66	**0.88**	0.47	[0.49]	0.65
	CCNN-IT[34]	0.75	0.69	0.86	0.40	0.45	0.63
	OR-CNN[35]	0 .71	0.63	[0.87]	0.41	0.31	0.58
	2DC[36]	[0.76]	0.71	0.85	0.45	**0.53**	[0.66]
	VGG16[37]	0.63	0.61	0.73	0.25	0.31	0.51
	Joint [38]	**0.79**	[0.80]	0.86	[0.54]	0.43	**0.68**
	Ours	0.54	**0.88**	0.77	**0.70**	0.33	0.64
MAE	VGP-AE [33]	0.82	1.28	0.70	1.43	**0.77**	1.00
	CCNN-IT[34]	1.23	1.69	0.98	2.72	1.17	1.57
	OR-CNN[35]	0.88	1.12	0.68	1.52	0.93	1.02
	2DC[36]	[0.75]	1.02	0.66	[1.44]	0.88	0.95
	VGG16[37]	0.93	1.04	0.91	1.51	1.10	1.10
	Joint [38]	0.77	[0.92]	[0.65]	1.57	**0.77**	[0.94]
	Ours	**0.48**	**0.47**	**0.64**	**0.67**	0.99	**0.65**

Table 5. ICC(3,1) and MAE scores in terms of 12 AUs are reported for the proposed method and the state-of-the-art methods on DISFA dataset. Bold numbers indicate the best performance; bracketed numbers indicate the second best.

	Method	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	Avg.
ICC	VGP-AE[33]	0.37	0.32	0.43	0.17	0.45	0.52	0.76	0.04	0.21	0.08	0.80	0.51	0.39
	CCNN-IT [34]	0.18	0.15	0.61	0.07	**0.65**	[0.55]	[0.82]	**0.44**	[0.37]	**0.28**	0.77	[0.54]	0.45
	OR-CNN [35]	0.33	0.31	0.32	0.16	0.32	0.28	0.71	[0.33]	**0.44**	[0.27]	0.51	0.36	0.36
	LT-all [39]	0.32	0.37	0.41	0.18	0.46	0.23	0.73	0.07	0.23	0.09	0.80	0.39	0.36
	2DC[36]	**0.70**	[0.55]	[0.69]	0.05	[0.59]	**0.57**	**0.88**	0.32	0.10	0.08	**0.90**	0.50	[0.50]
	BORMIR[40]	0.19	0.24	0.30	0.17	0.38	0.18	0.58	0.15	0.22	0.08	0.70	0.14	0.28
	KJRE [41]	0.27	0.35	0.25	0.33	0.51	0.31	0.67	0.14	0.17	0.20	0.74	0.25	0.35
	CFLF[42]	0.26	0.19	0.45	[0.35]	0.51	0.35	0.70	0.18	0.34	0.20	0.81	0.51	0.40
	Ours	[0.59]	**0.63**	**0.73**	**0.82**	0.49	0.50	0.73	0.29	0.21	0.03	**0.90**	**0.60**	**0.54**
MAE	VGP-AE[33]	1.02	1.13	0.92	0.10	0.67	[0.19]	0.33	0.46	0.58	0.19	0.69	0.65	0.57
	CCNN-IT[34]	0.87	0.63	0.86	0.26	0.73	0.57	0.55	0.38	0.57	0.45	0.81	0.64	0.61
	OR-CNN[35]	0.41	0.44	0.91	0.12	0.42	0.33	0.31	0.42	0.35	0.27	0.71	0.51	0.43
	LT-all [39]	0.44	0.39	0.96	[0.07]	0.41	0.31	0.40	[0.17]	[0.33]	[0.16]	0.61	0.46	0.39
	2DC[36]	[0.32]	0.39	[0.53]	0.26	0.43	0.30	[0.25]	0.27	0.61	0.18	[0.37]	0.55	0.37
	BORMIR[40]	0.87	0.78	1.24	0.58	0.76	0.77	0.75	0.56	0.71	0.62	0.89	0.87	0.78
	KJRE [41]	1.02	0.92	1.86	0.70	0.79	0.87	0.77	0.60	0.80	0.72	0.96	0.94	0.91
	CFLF[42]	[0.32]	[0.28]	0.60	0.12	[0.35]	0.27	0.42	0.18	**0.29**	[0.16]	0.53	**0.39**	[0.32]
	Ours	**0.16**	**0.08**	**0.40**	**0.02**	**0.23**	**0.12**	**0.22**	**0.14**	0.48	**0.12**	**0.27**	**0.39**	**0.22**

Table 6. F1 scores in terms of 12 AUs are reported for the proposed method on the BP4D dataset. Colored AU is removed during training, and only used for testing.

AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	***Avg***
44.1	69.0	73.8	80.2	62.2	88.1	89.2	38.5	53.7	60.6	27.6	68.2	62.9
89.9	77.5	53.6	66.9	39.7	70.4	81.5	52.2	86.6	45.7	63.8	37.3	63.6

the highest F1-score (65.5%), demonstrating the effectiveness of our proposed method.

Second, to show the effectiveness of Eq.(4) and Eq.(5), we set $scale = 1, bias = 0$ in Eq.(4). Since then RE-Net is equivalent to the vanilla Resnet18, by comparing our method and the vanilla Resnet18, we can see a **5.9%** improvement in F1-score, which demonstrates the effectiveness of Eq.(4,5) in improving the performance of AU recognition.

Third, the Pearson correlation coefficient is computed in each dataset and fixed in Eq.(5), which could be biased, as different datasets vary in subjects and tasks. To further investigate this issue, we try to learn the AU relation along with AU detection by setting the AU relation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as learnable parameters. Specifically, the $\omega_{i,j}$ of AU relation $\mathcal{G}$ is first randomly initialized, and then updated through training by minimizing the loss function in Eq.(6). As indicated by ***Ours+learnable*** in Table.7, a 63.7% F1-score is achieved. Although the performance is not as good as ours with fixed AU relation, its performance is **4.1%** higher than the Resnet18 baseline, and outperforms the state-of-the-art methods. More importantly, it has the potential to learn a general dataset-independent AU relation.

Table 7. F1-scores of different methods with AU relation are reported on the BP4D dataset. *Learnable indicates the AU relation is learned through training rather than a fixed factor.*

Method	AU relation applied to:	***Avg***
JPML [10]	*image patch*	45.9
DRML [11]	*image patch*	48.3
DSIN [12]	*deep features*	58.9
SRERL[13]	*deep features*	62.9
ResNet18	*none*	59.6
Ours	*model's parameters*	**65.5**
Ours+ *learnable*	*model's parameters*	[63.7]

Imbalance issue: Most AU databases are imbalanced. Take BP4D as example, the occurrence rate of AU6, AU7, AU10, AU12 and AU14 are almost 2-3 times more than the others. To deal with the imbalanced issue, most works apply a data augmentation method (*i.e.,*duplication) to increase the frames of AUs with less occurrence rate. To evaluate the impact of data augmentation, we duplicate the AUs (AU1, AU2, AU15, AU23 and AU24) one time if a positive label is observed in the training dataset. As we can see in Table.8, ResNet18 achieves 1.3% improvement by using the augmented training dataset; while our method shows only 0.1% difference with/without data augmentation, indicating the effectiveness of our proposed method in handling the data imbalance issue.

Table 8. F1-score of ablation study for 12 AUs on the BP4D database. DA: *data augmentation*.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	***Total***
ResNet18 w/o DA	44.5	45.1	51.1	81.2	76.8	87.6	86.8	67.9	44.2	57.5	42.0	30.8	59.6
ResNet18 + DA	49.4	53.0	58.0	79.2	72.7	86.0	89.6	68.1	34.4	65.0	42.3	33.5	60.9
Ours + DA	61.9	50.3	62.3	80.0	73.7	87.3	90.3	70.4	42.0	67.4	47.3	55.5	65.6
Ours w/o DA	57.7	59.0	66.9	76.3	77.0	88.9	89.8	70.9	42.0	62.8	44.8	49.3	65.5

Recognize New AU: Inspired by recent works[17][18][19][16] that extend batch-normalization to address domain adaptation tasks, we conduct two initial experiments to verify the ability of recognizing new AU. In Table.6, the label of the colored AU is selected to remove during training. During testing, the AU-specific parameters (γ, β) are calculated by using Eq.(4) and Eq.(5). As we can see in Table.6, transferring the AU-specific parameters to the new AU, our method shows comparable results in recognizing the unseen AU23 and AU24.

Feature Visualization: To provide insight into the feature space for individual AUs, we first extract the features for the testing images on the BP4D dataset using our proposed method and the ResNet18. Fig.3 shows the t-SNE [43] embedding of frames, which are colored in terms of AU4, AU10, AU12 and AU24 (different colors means presence or absence of a specific AU). ResNet18 extracts a single representation for each input image for multiple AUs detection, hence the shapes of t-SNE embedding are all the same. On the contrary, our method extracts different features for individual AU, therefore the shapes of t-SNE embedding are different in AU4, AU10, AU12 and AU24. By comparing the AU related projection, we may find that the features extracted for individual AU by our method are more robust than the features extracted by ResNet18, for example, the green points of AU4 and AU24, which are more challenging to recognize than AU10 and AU12, are more concentrated in our method than ResNet18.

4.7 Conclusion

In this paper, we format the multi-label AU detection problem as a domain adaptation problem, and propose a new AU relationship embedded deep model (**RE-Net**) for AU detection, which contains both shared and AU-specific parameters. The AU relation is modeled in the model's AU-specific parameters space, therefore, the deep model can be optimized effectively for individual AU. We also apply a new training strategy that will not add extra burden for the model training. Extensive experiments are conducted on the widely used datasets for both AU recognition and AU intensity estimation, demonstrating the superiority of the proposed method over the state-of-the-art methods.

One concern of the proposed method is the running efficiency, as the model needs to be run multiple times for detecting different AUs. We measure the inference time of our method with 12 AUs on the NVIDIA GeForce 2080Ti

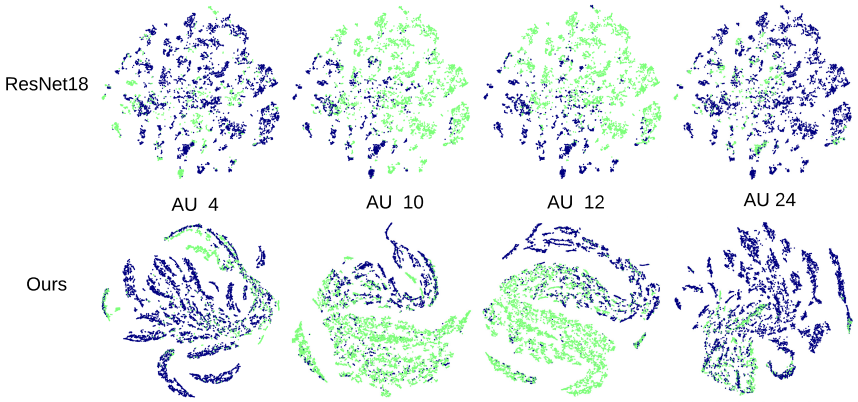

Fig. 3. A visualization of t-SNE embedding using deep features on the BP4D database by coloring each frame from testing images in terms of AU4, AU10, AU12 and AU24 (blue and green means the absent and occurrence of individual AU respectively). Best viewed in color.

GPU, which is around 5.4ms/image, equivalent to 183 FPS (frame per second). Although the processing speed is much slower than the ResNet18 baseline, our method has achieved a much higher performance, and the 183 FPS processing speed is more than enough for the real-time processing requirement.

Our future work is to improve the model's efficiency as well as to extend it for detection of new AUs.

5 Acknowledgment

The material is based on the work supported in part by the NSF under grant CNS-1629898 and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

References

1. Friesen, E., Ekman, P.: Facial action coding system: a technique for the measurement of facial movement. Palo Alto **3** (1978)
2. Li, Y., Chen, J., Zhao, Y., Ji, Q.: Data-free prior model for facial action unit recognition. IEEE Transactions on affective computing **4** (2013) 127–141
3. Wang, Z., Li, Y., Wang, S., Ji, Q.: Capturing global semantic relationships for facial action unit recognition. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 3304–3311
4. Tong, Y., Ji, Q.: Learning bayesian networks with qualitative constraints. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2008) 1–8

5. Liao, W., Ji, Q.: Learning bayesian network parameters under incomplete data with domain knowledge. Pattern Recognition **42** (2009) 3046–3056
6. Benitez-Quiroz, C.F., Wang, Y., Martinez, A.M.: Recognition of action units in the wild with deep nets and a new global-local loss. In: ICCV. (2017) 3990–3999
7. Zhang, Y., Dong, W., Hu, B.G., Ji, Q.: Classifier learning with prior probabilities for facial action unit recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5108–5116
8. Peng, G., Wang, S.: Weakly supervised facial action unit recognition through adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2188–2196
9. Peng, G., Wang, S.: Dual semi-supervised learning for facial action unit recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8827–8834
10. Zhao, K., Chu, W.S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2207–2216
11. Zhao, K., Chu, W.S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3391–3399
12. Corneanu, C., Madadi, M., Escalera, S.: Deep structure inference network for facial action unit recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 298–313
13. Li, G., Zhu, X., Zeng, Y., Wang, Q., Lin, L.: Semantic relationships guided representation learning for facial action unit recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8594–8601
14. Shao, Z., Zou, L., Cai, J., Wu, Y., Ma, L.: Spatio-temporal relation and attention learning for facial action unit detection. arXiv preprint arXiv:2001.01168 (2020)
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML (2015)
16. Mancini, M., Bulo, S.R., Caputo, B., Ricci, E.: Adagraph: Unifying predictive and continuous domain adaptation through graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6568–6577
17. Sun, Q., Liu, Y., Chua, T.S., Schiele, B.: Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 403–412
18. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7354–7362
19. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. Pattern Recognition **80** (2018) 109–117
20. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32** (2014) 692–706
21. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing **4** (2013) 151–160
22. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J.F., Ji, Q., Yin, L.: Multimodal spontaneous emotion corpus for human behavior analysis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)

23. Li, W., Abtahi, F., Zhu, Z., Yin, L.: Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. arXiv preprint arXiv:1702.02925 (2017)
24. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86** (1979) 420
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
26. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. (2019) 8024–8035
27. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of machine learning research **9** (2008) 1871–1874
28. Shao, Z., Liu, Z., Cai, J., Ma, L.: Deep adaptive attention for joint facial action unit detection and face alignment. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 705–720
29. Yang, H., Yin, L.: Learning temporal information from a single image for au detection. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE (2019) 1–8
30. Niu, X., Han, H., Yang, S., Huang, Y., Shan, S.: Local relationship learning with person-specific shape regularization for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 11917–11926
31. Mei, C., Jiang, F., Shen, R., Hu, Q.: Region and temporal dependency fusion for multi-label action unit detection. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE (2018) 848–853
32. Ma, C., Chen, L., Yong, J.: Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. Neurocomputing **355** (2019) 35–47
33. Eleftheriadis, S., Rudovic, O., Deisenroth, M.P., Pantic, M.: Variational gaussian process auto-encoder for ordinal prediction of facial action units. In: Asian Conference on Computer Vision, Springer (2016) 154–170
34. Walecki, R., Pavlovic, V., Schuller, B., Pantic, M., et al.: Deep structured learning for facial action unit intensity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 3405–3414
35. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 4920–4928
36. Linh Tran, D., Walecki, R., Eleftheriadis, S., Schuller, B., Pantic, M., et al.: Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3190–3199
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
38. Sánchez-Lozano, E., Tzimiropoulos, G., Valstar, M.: Joint action unit localisation and intensity estimation through heatmap regression. BMVC (2018)
39. Kaltwang, S., Todorovic, S., Pantic, M.: Latent trees for estimating intensity of facial action units. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 296–304

40. Zhang, Y., Zhao, R., Dong, W., Hu, B.G., Ji, Q.: Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7034–7043
41. Zhang, Y., Wu, B., Dong, W., Li, Z., Liu, W., Hu, B.G., Ji, Q.: Joint representation and estimator learning for facial action unit intensity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3457–3466
42. Zhang, Y., Jiang, H., Wu, B., Fan, Y., Ji, Q.: Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 733–742
43. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9** (2008) 2579–2605