Emotion-aware Contrastive Learning for Facial Action Unit Detection

Xuran Sun[1,2], Jiabei Zeng[1], Shiguang Shan[1,2,3]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Peng Cheng Laboratory, Shenzhen, 518055, China

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9666945

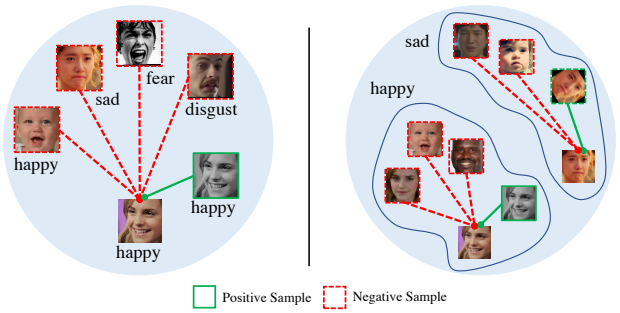
***Abstract*— Current AU datasets lack sufficiency and diversity because annotating facial action units (AUs) is laborious. The lack of labeled AU datasets bottlenecks the training of a discriminative AU detector. Compared with AUs, the basic emotional categories are relatively easy to annotate and they are highly correlated to AUs. To this end, we propose an Emotion-aware Contrastive Learning (EmoCo) framework to obtain representations that retain enough AU-related information. EmoCo leverages enormous and diverse facial images without AU annotations while labeled with the six universal facial expressions. EmoCo extends the prevalent self-supervised learning architecture of Momentum Contrast by simultaneously classifying the learned features into different emotional categories and distinguishing features within each emotional category in instance level. In the experiments, we train EmoCo using AffectNet dataset labeled with emotional categories. The EmoCo-learned features outperform other self-supervised learned representations in AU detection tasks on DISFA, BP4D, and GFT datasets. The EmoCo-pretrained models that fine-tuned on the AU datasets outperform most of the state-of-the-art AU detection methods.**

I. INTRODUCTION

Automatic facial behavior analysis has attracted increasing attention due to its wide applications in human-computer interaction. To study facial behavior comprehensively, Ekman and his colleagues proposed Facial Action Unit System (FACS) [8] to objectively characterize facial actions. FACS defines a unique set of about 40 discernible and non-overlapping facial movements, which are called Action Units (AUs). Automatic AU detection has become one of the most significant domains of facial behavior analysis as it is very promising in a wide range of applications, such as affect analysis and mental health assessment. Recently, the application of deep supervised learning methods provides new approaches for representation learning and feature extraction [18], [19], [37], [29], which tremendously promotes the performance of AU detection.

However, these supervised AU detection methods deeply rely on a large amount of AU annotations while AU datasets in the literature are still constrained by the number of coded AUs, samples, and subjects [36], due to the demanding coding process. Usually, it takes 30 minutes or more for a specially trained coder to manually annotate an AU for a one-minute video.

This work is partially supported by National Key R&D Program of China (No. 2017YFA0700800) and National Natural Science Foundation of China (No. 62176248, 61702481).



(a) Conventional Contrastive Learning (b) Emotion-aware Contrastive Learning

Fig. 1. Conventional vs emotion-aware contrastive learning. They both encourage the positive samples to be close and the negative samples to be far away from each other. The positive samples are two random augmentations of an instance. (a) In conventional contrastive learning, the negative samples are different instances in various emotional categories. (b) In emotion-aware contrastive learning, the negative samples are different instances within the same emotional category as the positive ones.

Annotating universal facial expressions is much easier than annotating AUs, and AUs are highly relevant to universal facial expressions according to the psychological studies [15]. As discovered by psychologists, the understanding of six prototypical facial expressions, i.e., anger, happiness, fear, surprise, sadness, and disgust, are relatively common across different cultures [7]. Therefore, the annotators for facial expressions could be less professional than those for AUs. Manually labeling facial expressions becomes efficient via the crowding source strategy, which contributes to the large-scale annotated facial expression datasets, e.g., AffectNet [24] and RAF-DB [17]. Besides, AUs depict finer facial behaviors than the prototypical facial expressions categories. Each facial expression can be characterized by a concrete combination of AUs. For instance, the "anger" usually occurs with the combination of AU4, AU5, and AU24.

To learn discriminative AU representations leveraging large amount emotion-annotated images without AU labels, we propose the Emotion-aware Contrastive Learning (EmoCo) framework. EmoCo treats the prototypical facial expressions as coarse categories and treats each image with different AU combinations as a fine subclass within every category. Inspired by the coarse-to-fine framework in [1], EmoCo extends the prevalent self-supervised learning architecture of Momentum Contrast (MoCo) [11] by encouraging the learned features to be classified into one of the emotional

categories and to be instance-distinguishable within each emotional category. Fig. 1 illustrates the key differences between the emotion-aware contrastive learning in EmoCo and the conventional one in MoCo [11]. They both encourage the positive samples to be close and the negative samples to be far away from each other. The positive samples are two random augmentations of an instance. In conventional contrastive learning, the negative samples are from different instances in various emotional categories. The conventional methods are hard to learn AU features because it may focus on emotional-unrelated information, e.g., background, identities, age, gender. While in emotion-aware contrastive learning, to keep only the emotion-related information, the learned features are classified into discrete emotional classes. Simultaneously, the negative samples and the positive ones within the same emotional category are contrasted. Thus, the EmoCo-learned features implicitly capture the differences among AUs combinations in different prototypical facial expressions and are capable in representing fine-grained facial behaviors.

The contributions of this work can be summarized as :

1. We leverage the emotion-annotated images to learn discriminative AU representations, which alleviates the demands for adequate manual AU annotations.

2. We propose the Emotion-aware Contrastive Learning (EmoCo) framework. It regards AUs as finer descriptors of facial behaviors than emotional categories and then adopts a coarse-to-fine contrastive learning paradigm [1].

3. We conduct extensive experiments and validate the advantages of the EmoCo-learned features and the finetuned EmoCo models over other representations or state-of-the-art supervised AU detection methods.

II. RELATED WORK

A. Facial Action Unit Detection

AU detection has been studied for decades and various methods have been proposed. According to the composition of training data, present AU detection methods can be categorized into following groups:

Fully supervised methods construct AU detectors by learning from training examples with complete AU annotations indicating ground truth. Zhao et al. [37] propose a locally connected convolutional layer that learns region-specific AU representations. EAC-Net [19] extracts features around facial landmarks that are robust with respect to non-rigid shape changes. JAA-Net [29] jointly estimates the location of landmarks and the presence of action units. Li et al. [16] and Corneanu et al. [6] incorporate graphical models in their proposed frameworks for AU relationship reasoning and modeling. These methods have achieved promising performance on AU annotated datasets, e.g., DISFA [23], BP4D [34]. However, they are overly dependent on the annotated training data and lacking in generalizability.

To alleviate the demand for enormous and accurate AU annotations, researchers try to use data with noisy, incomplete or none labels to learn AU representations. Previous works for weakly supervised AU detection mainly focused on utilizing face images with incomplete labels or noisy labels to improve the AU detection accuracy. Wu et al. [33] proposed to use Restricted Boltzmann Machine to model the AU distribution, which is further used to train the AU classifiers with partially labeled data. In [36], Zhao et al. propose a weakly supervised clustering method for pruning noise labels and train the AU classifiers with re-annotated data. Semi-supervised methods leverage both labeled and unlabeled data for AU detection. Niu et al. [25] propose a novel multi-label co-regularization method for semi-supervised AU recognition. Self-supervised methods adopt pseudo labels, which are inferred from the structure of the unlabeled data itself, as supervisory signals to learn AU representations. Wiles et al. [32] and Li et al. [21] transform the source frame to the target frame of the same person according to the decoded displacement field from the learned features. Lu et al. [22] use a triplet-based ranking approach that learns to rank the frames based on their temporal distance from an anchor frame.

Some other works propose to recognize AUs utilizing training data from other related tasks or in a multi-task manner. Peng et al. [26] utilize the prior knowledge of AUs and emotions to generate pseudo AU labels for training from facial images with only emotion labels. Zhang et al. [35] propose a knowledge-driven strategy for jointly training multiple AU classifiers without any AU annotation by leveraging prior probabilities on AUs. Results in [31], [27] prove that the AU detection and facial expression recognition can promote each other in the multi-task learning paradigm. Our proposed EmoCo learns AU representations in a multi-task manner. It focuses on separating emotional categories and learning fine-grained AU features at the same time.

B. Contrastive Learning

Contrastive learning is a machine learning technique used to learn the general features of data by teaching the model which data points are similar or different. It can be applied under both unsupervised and supervised settings.

When working with unlabeled data, contrastive learning is one of the most powerful approaches in self-supervised learning. Self-supervised contrastive learning models have achieved an amazing performance even comparable to supervised models attributed to its ability of instance feature discrimination. SimCLR [4] defines positive pairs as two augmentations of the same image and contrasts them with other images of the same batch. MoCo [11] builds the positive pairs in the same way, yet contrasts them with negative samples stored in a dynamic dictionary produced by a encoder with moving average parameters. SWAV [3] contrasts the cluster assignments of different views instead of contrasting features directly. BYOL [10] discards negative samples and directly computes the similarity between positive pairs by introducing a slowly progressing momentum encoder and a prediction head. On the basis of BYOL [10], SimSiam [5] abandons momentum update and brings in gradient stop.

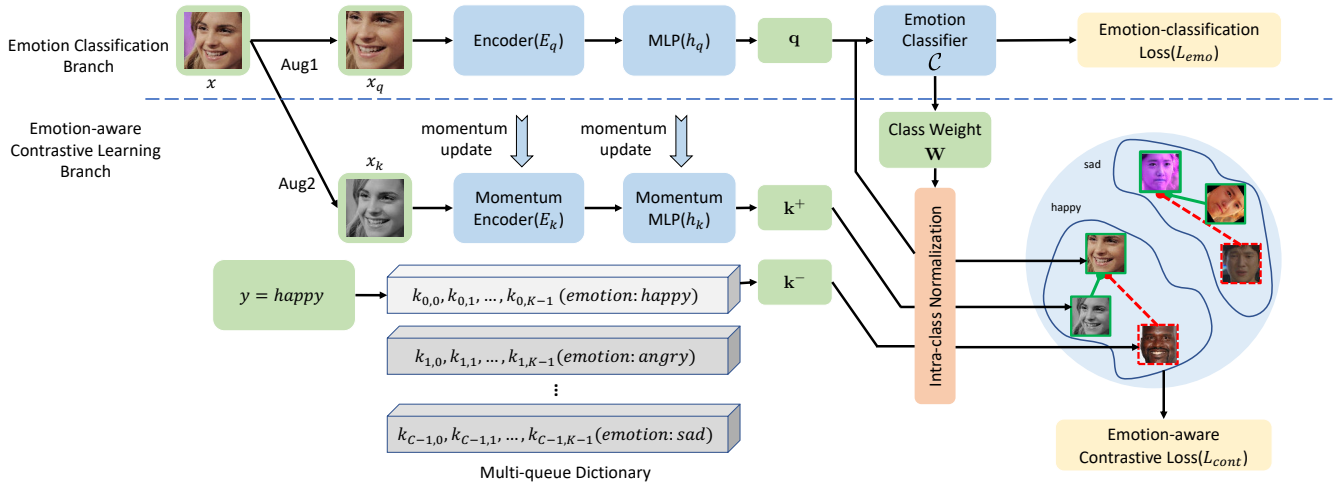


Fig. 2. EmoCo's framework overview. Given an input image x, in the emotion classification branch, we encode its augmentation x_q into feature $\mathbf{q}$, and then train an emotion classifier $\mathcal{C}$ to remove non-emotional information in the features. In the emotion-aware contrastive learning branch, we get a positive feature $\mathbf{k}^+$ by encoding another augmentation x_k of the input x. The negative features $\mathbf{k}^-$ are taken from a multi-queue dictionary according to the emotional label y of the input image. Each queue of the dictionary stores historical keys from the same emotion. Then, $\mathbf{q}$, $\mathbf{k}^+$, and $\mathbf{k}^-$ are intra-class normalized into an emotion-specific space. The contrastive loss forces the positive pair $(\mathbf{q}, \mathbf{k}^+)$ to be close and the negative pair $(\mathbf{q}, \mathbf{k}^-)$ to be far away.

When working with labeled data, contrastive learning can still show its effectiveness. CLIP [28] jointly trains a text encoder and an image feature extractor over a contrastive learning task that predicts which caption goes with which image. Supervised Contrastive Loss [13] aims to leverage label information more effectively than cross entropy, imposing that normalized embeddings from the same class are closer together than embeddings from different classes. ANCOR [1] uses contrastive learning to learn features of the query set in a few-shot learning setting.

III. Method

A. Overview

To learn discriminative AU representations without AU annotations, we propose emotion-aware contrastive learning (EmoCo) that adopts a coarse-to-fine contrastive learning framework [1] leveraging expression-labeled images. EmoCo regards each emotional category as a coarse class and the instances with different AU combinations as fine subclasses, considering the fact that AUs are finer descriptors of facial behaviors than emotional categories.

Fig. 2 illustrates the outline of EmoCo. EmoCo learns the features by extending the learning framework of MoCo [11], which is a prevalent self-supervised learning framework that learns instance-distinguishable features. Unlike MoCo [11], EmoCo categorizes the features into different coarse emotion classes and meanwhile encourages the learned features to be distinguishable in instance level within each coarse class. As can be seen in Fig. 2, EmoCo consists of two branches, i.e., an emotion classification branch (upper) and an emotion-aware contrastive learning branch (lower). In the emotion classification branch, the learned features are optimized under facial expression supervision. Therefore, the features tend to represent the facial expression other than irrelevant factors, e.g., the identities and ages. In the emotion-aware contrastive learning branch, EmoCo learns fine-grained AU features that can separate different instances within each emotional category. By simultaneously minimizing the emotion-classifcation loss $\mathcal{L}_{emo}$ and the emotion-aware contrastive loss $\mathcal{L}_{cont}$ in the two branches, EmoCo is trained in an end-to-end manner by

$$\min(\lambda \mathcal{L}_{emo} + \mathcal{L}_{cont}), \tag{1}$$

where λ is the coefficient that balances the importance of the emotion classification and emotion-aware contrastive learning, whose details are presented below.

B. Emotion Classification

The emotion classification branch requires the learned features to be categorized into one of the prototypical facial expressions under facial expression supervision, which provides a coarse guidance for further AU representations learning.

As shown in Fig. 2, given a training image x and its emotion label $y \in \{0, 1, 2, \cdots, C-1\}$, where C is the number of emotional categories, we randomly augment x to get x_q, and then encode x_q to obtain a 128-dim embedding $\mathbf{q}$ with the online encoder E_q and the multiple layer perceptron(MLP) h_q. E_q is a CNN network with global average pooling on the top and h_q includes L_2 nomalization of its output. Thus, $\mathbf{q}$ is a unit vector. We use $\mathbf{q}$ to train the emotion classifier $\mathcal{C}$, a single-layer fc, to predict the emotional category of the input image x. The parameters of the encoder E_q, MLP h_q, and emotion classifier $\mathcal{C}$ are optimized by minimizing Emotion-classification Loss $\mathcal{L}_{emo}$:

$$\mathcal{L}_{emo} = -\log \frac{\exp(\mathbf{w}_y \mathbf{q})}{\sum_{j=0}^{C-1} \exp(\mathbf{w}_j \mathbf{q})}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{C \times 128}$ is the weight matrix of the emotion classifier $\mathcal{C}$. $\mathbf{w}_y$ is the y-th row of $\mathbf{W}$ and denotes the parameters of the classifier $\mathcal{C}$ for the y-th emotional category.

The emotion-classification loss L_{emo} is minimized when $\mathbf{w}_y \cdot \mathbf{q}$ is maximized and $\mathbf{w}_{j \neq y} \cdot \mathbf{q}$ is minimized. This happens

when $\mathbf{q}$ shares the identical direction with $\mathbf{w}_y$. And this is the same for all images belonging to the emotion class y. We could regard $\tilde{\mathbf{w}}_y = \frac{\mathbf{w}_y}{||\mathbf{w}_y||}$, the L_2 normalization of $\mathbf{w}_y$, as the prototype of the y-th emotional category. By minimizing $\mathcal{L}_{emo}$, the emotion classification branch forces the learned features to gather around their corresponding prototype $\tilde{\mathbf{w}}_y$ and separates the whole feature space into several discrete emotional categories, offering a coarse supervision to remove the emotion-unrelated information in the features.

C. Emotion-aware Contrastive Learning

Although the emotion classification branch separates the learned features into the coarse emotional categories, it neglects the diversity of facial behaviors in each emotion. Distinguishing the fine-grained facial behaviors is essential to represent the AUs. To learn fine-grained features, EmoCo encourages the learned features to be instance-distinguishable in each emotional category by enlarging the distance between the features of two instances. However, the features of different instances within each emotional category are forced to be close in the emotion classification branch. It is conflicting to simultaneously require the features to be far away and to be close. To this end, we propose emotion-aware contrastive learning with intra-class normalization [1].

During training, the emotion-aware contrastive learning branch chooses the positive samples as the two augmentations of the same instance. The negative samples are different instances from the same emotional category. Then, the intra-class normalized positive samples are forced to be close and the normalized negative ones are forced to be apart. Below, we will introduce the process of sample selection and intra-class normalization [1] in detail.

1) Positive Samples: As shown in Fig. 2, given an input image x, we randomly augment x twice to get its two views x_q and x_k to form the positive pair. x_q is encoded by the online encoder E_q and embedded by an MLP projection head h_q to get the positive query $\mathbf{q}$. Correspondingly, we pass x_k through the momentum encoder E_k and momentum MLP h_k to compute the positive key $\mathbf{k}^+$. E_k and h_k have the same structure as their counterpart E_q and h_q. Formally, denoting the parameters of the sequential $E_k \rightarrow h_k$ as θ_k, and those of $E_q \rightarrow h_q$ as θ_q, we update θ_k by:

$$\theta_k := m\theta_k + (1-m)\theta_q. \tag{3}$$

Here $m \in [0, 1)$ is a momentum coefficient. θ_q is updated by back-propagation.

2) Negative Samples: For query $\mathbf{q}$, its negative sample is selected from the historical samples that belong to the same emotional category as itself. EmoCo stores the candidate negative samples by maintaining a dictionary to save their encoded featured from the preceding batches. In order to store negative samples from different emotion classes separately, we extend the single-queue dictionary in MoCo [11] to a multi-queue dictionary. As shown in Fig. 2, samples from the different emotional categories are stored in different queues. The stored keys in the dictionary are progressively replaced. The samples in current input mini-batch are enqueued according to their emotional category, and the same quantity of oldest samples are popped. The size of each queue in the dictionary is K, which is a hyper-parameter and is much larger than the batch size.

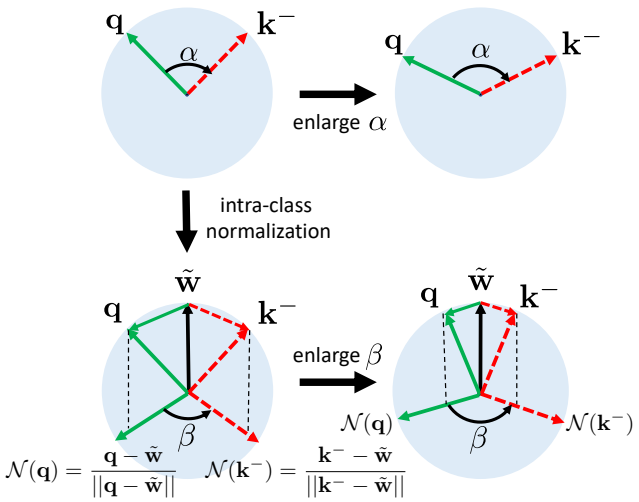


Fig. 3. L_2 normalization vs intra-class normalization. **Top:** After L_2 normalized, minimizing contrastive loss $\mathcal{L}_{cont}$ will enlarge the angle α between the negative pair $(\mathbf{q}, \mathbf{k}^-)$ of the same emotion class. **Bottom:** After intra-class normalized, minimizing contrastive loss $\mathcal{L}_{cont}$ will enlarge the angle β between the negative pair $(\mathcal{N}(\mathbf{q}), \mathcal{N}(\mathbf{k}^-))$ in the emotion-sepcific space and push $(\mathbf{q}, \mathbf{k}^-)$ towards the prototype $\tilde{\mathbf{w}}$ in the original space.

3) Intra-class Normalization: EmoCo requires the positive samples to be close and the negative samples to be faraway. However, since the negative samples are from the same emotional category, they are also forced to be close in emotion classification by (2). To relieve the conflict, we introduce the intra-class normalization [1] to transform the features $\mathbf{q}$ into a class-centered space as

$$\mathcal{N}(\mathbf{q}) = \frac{\mathbf{q} - \tilde{\mathbf{w}}}{||\mathbf{q} - \tilde{\mathbf{w}}||}, \tag{4}$$

where $\tilde{\mathbf{w}}$ is an L_2-normalized weight vector corresponding to one row of the emotion classifier $\mathcal{C}$'s weight matrix $\mathbf{W}$.We can regard $\tilde{\mathbf{w}}$ as the center or the prototype of an emotion class. Then, we minimize the emotion-aware contrastive loss $\mathcal{L}_{cont}$ in the intra-class normalized space as

$$\mathcal{L}_{cont} = -\log \frac{e^{\langle \mathcal{N}(\mathbf{q}), \mathcal{N}(\mathbf{k}^+) \rangle / \tau}}{e^{\langle \mathcal{N}(\mathbf{q}), \mathcal{N}(\mathbf{k}^+) \rangle / \tau} + \sum_{i=0}^{K-1} e^{\langle \mathcal{N}(\mathbf{q}), \mathcal{N}(\mathbf{k}_i^-) \rangle / \tau}}, \tag{5}$$

where $\langle , \rangle$ represents the inner product of two vectors, and τ is the temperature.

Fig. 3 illustrates the advantage of intra-class normalization over conventional L_2 normalization. In L_2 normalization (top left), the query $\mathbf{q}$ and its negative sample $\mathbf{k}^-$ are normalized to unit vectors. By minimizing the contrastive loss, the angle α between $\mathbf{q}$ and $\mathbf{k}$ is enlarged (top right). However, this is in direct conflict with the interest of emotion classification loss $\mathcal{L}_{emo}$ that tries to pull $\mathbf{q}$ and $\mathbf{k}^-$ towards their prototype $\tilde{\mathbf{w}}$. In intra-class normalization(bottom left), by minimizing the contrastive loss, the angle β between the intra-class normalized feature $\mathcal{N}(\mathbf{q})$and $\mathcal{N}(\mathbf{k}^-)$ is enlarged while $\mathbf{q}$ and

$\mathbf{k}$ move towards their class prototype $\tilde{\mathbf{w}}$, unifying the effects of $\mathcal{L}_{cont}$ and $\mathcal{L}_{emo}$ at the same time.

IV. Experiments

In this section, we validate the effectiveness of the proposed EmoCo. First, we compare EmoCo with other state-of-the-art AU detection methods. Then, we analyze the components and parameters.

A. Datasets

We pretrain EmoCo on a facial expression labeled dataset——AffectNet, and provide evaluations of the pretrained model on three widely used datasets, i.e., GFT [9], DISFA [23], and BP4D [34].

1) AffectNet: AffectNet [24] is by far the largest database of 7 facial expressions (including neutral) in the wild. We utilize about 280,000 images which are annotated with seven discrete facial expressions to pretrain EmoCo.

2) GFT: This dataset contains 96 participants in 32 three-person groups. It has pre-divided the training set and the test set. The moderate out-of-plane head motion and occlusion in this dataset make AU detection challenging. We follow the original train/test splits in [9] (about 108000 facial images for training and 24600 images for valuation) and use 10 AUs for evaluation. We perform three-time test on its validation set and report the average performance to reduce the bias.

3) DISFA: It consists of 26 participants, whose AUs are labeled with intensities from 0 to 5. In each frame, AUs with intensities greater than 1 are considered as positive, while the others are annotated as negative. Totally, about 130,000 AU-labeled frames are obtained. We split the dataset into 3 folds based on subject IDs and conduct 3-fold cross-validation to evaluate model performance.

4) BP4D: It is a spontaneous facial AU dataset containing 328 videos from 41 subjects (23 females and 18 males). Each subject is involved in 8 sessions, and their spontaneous facial expressions are recorded. 12 AUs are annotated for all the video frames, and there are about 140,000 images with AU labels. A 3-fold cross-validation is conducted on the dataset.

B. Implementation Details

1) Data Preprocessing: For all the images used in the experiments, we utilize an open source SeetaFace[1] face detector to detect the face rectangle and five facial landmarks. All of the face images are aligned and cropped to 256×256 based on the detected landmarks. During training, the data augmentation settings follows MoCo [11].

2) Optimizer: We use SGD for pretraining since EmoCo doesn't need a large-batch optimizer. We use a learning rate of $lr \times BatchSize/256$, with a base $lr = 0.03$. The learning rate has a cosine decay scheduler. The weight decay is 0.0001 and the SGD momentum is 0.9.

The batch size is 256 by default, which is friendly to typical 4-GPU implementations. We use batch normalization (BN) synchronized across devices, following [11].

[1]https://github.com/seetaface/SeetaFaceEngine

3) Network Structure: We use ResNet-50 [12] as the default backbone of the encoder. Our MLP projector consists of 2 fully connected layers. Its output dim is 128 and its hidden layer is 2048-d with a ReLU activation function. The emotion classifier $\mathcal{C}$ is a single-layer fc.

4) Hyper-parameters: Unless specified, the size of each queue in the dictionary is $K = 65536$, infoNCE temperature is $\tau = 0.2$, weight coefficient is $\lambda = 1$ for $\mathcal{L}_{emo}$, and momentum coefficient is $m = 0.999$ for the momentum encoder in our experiments. We perform 200-epoch pretraining in all experiments.

C. Evaluation Protocols

We pretrain EmoCo on the AffectNet [24] dataset, and evaluate the pretrained EmoCo's performance under both linear and finetuning protocol in each AU dataset.

Linear protocol. A linear protocol is to train a classifier using frozen representations learnt by the pretrained encoder on the training set of the target AU dataset and test the classifier's performance on the corresponding validation set.

Finetuning protocol. A finetuning protocol is to add an AU classifier on the top of the pretrained encoder, finetune them on the training set of the target AU dataset and test their performance on the corresponding validation set.

Following previous AU detection methods, we use F1 score as performance indicator for all the experiments. We also report the average F1 score of all AUs (denoted as Avg.).

D. Comparison with Other Methods

We compare EmoCo with the state-of-the-art supervised and self-supervised AU detection methods under both linear and finetuning protocol. Table I, II, III report the F1 score of these methods on GFT [9], DISFA [23], and BP4D [34].

Comparision with self-supervised methods: The EmoCo is compared with the state-of-the-art self-supervised methods: MoCo [11], Fab-Net [32], TAE [20], and Temporal Ranking [22] under both linear and finetuning protocol.

Under linear protocol, the average F1-scores of AU classifier trained with frozen EmoCo-encoded features surpass nearly all the self-supervised methods on three datasets (except 1% reduction than TAE [20] on BP4D [34] dataset). By combining auxiliary expression supervision and contrasitve learning, EmoCo is able to extract more expressive features and find more subtle differences of AUs than traditional self-supervised methods.

Under finetuning protocol, EmoCo outperforms all the listed self-supervised methods on three AU datasets. And the advantage of EmoCo is the most apparent on GFT [9] dataset, which is the most challenging dataset of the three. The results demonstrate EmoCo's generalizability and stability when working as pretrained model to be finetuned on downstream AU detection tasks.

Comparision with supervised methods: We compare EmoCo with the state-of-the-art supervised AU detection methods, including DRML [37], EAC [19], JAA [29], DSIN [6], and SRERL [16]. For fairness, we only make comparison under finetuning protocol.

TABLE I

F1 SCORES (IN %) OF 10 AUS BY THE PROPOSED EMOCO AND THE STATE-OF-THE-ART METHODS ON GFT DATASET. * MEANS THAT THE RESULTS ARE REPORTED IN THE ORIGINAL PAPERS.

	Methods/AU	1	2	4	6	10	12	14	15	23	24	Avg.
Supervised	AlexNet [14]*	44.0	46.0	2.0	73.0	72.0	82.0	5.0	19.0	43.0	42.0	42.8
	ResNet-50 [12]	23.5	37.8	3.5	79.1	70.1	82.1	20.9	11.7	49.1	40.3	41.8
Self-supervised	MoCo [11]	21.7	38.1	10.2	74.7	79.1	80.9	**25.9**	30.5	49.3	45.2	45.6
	MoCo(finetune) [11]	45.3	48.2	20.3	80.7	78.8	78.1	22.6	46.0	53.9	50.3	52.4
	Fab-Net [32]	44.4	42.3	9.4	60.6	68.7	70.4	8.7	1.7	5.5	20.8	33.3
	Fab-Net(finetune) [32]	33.3	52.6	12.4	80.2	75.6	82.7	16.6	37.1	46.1	46.6	48.3
	TAE [20]*	46.3	48.8	13.4	76.7	74.8	81.8	19.9	42.3	50.6	50.0	50.5
	TAE(finetune) [20]	30.5	46.4	20.0	77.7	**79.9**	83.0	18.9	44.5	47.9	47.5	49.6
	Temporal Ranking [22]	19.5	36.1	5.4	63.0	69.8	68.2	11.2	21.6	39.5	36.0	37.0
	Temporal Ranking(finetune) [22]	58.8	56.8	33.2	72.5	76.2	80.8	19.9	46.8	55.2	47.3	54.7
Ours	EmoCo	51.8	42.9	22.9	79.8	77.0	**85.2**	23.4	42.5	55.4	49.6	53.0
	EmoCo(finetune)	**65.9**	**55.9**	**40.7**	**83.1**	75.1	81.4	21.3	**48.5**	**58.0**	**56.5**	**58.6**

TABLE II

F1 SCORE (IN %) OF 8 AUS BY THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHODS ON THE DISFA DATASET. * MEANS THAT THE RESULTS ARE REPORTED IN THE ORIGINAL PAPERS.

	Methods/AU	1	2	4	6	9	12	25	26	Avg.
Supervised	DRML [37]*	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
	EAC [19]*	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
	JAA [29]*	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
	DSIN [6]*	42.4	39.0	**68.4**	28.6	46.8	70.8	**90.4**	42.2	53.6
	SRERL [16]*	**45.7**	**47.8**	59.6	47.1	45.6	73.5	84.3	43.6	55.9
Self-supervised	MoCo [11]	13.8	16.4	43.8	53.1	37.1	74.2	75.5	43.7	44.7
	MoCo(finetune) [11]	31.1	27.7	59.8	40.3	35.0	70.1	84.6	**60.8**	51.2
	Fab-Net [32]*	15.5	16.2	43.2	50.4	23.2	69.6	72.4	42.4	41.6
	Fab-Net(finetune) [32]	20.8	20.3	54.1	46.9	45.5	71.2	82.7	51.7	49.1
	TAE [20]*	21.4	19.6	64.5	46.8	44.0	73.2	85.1	55.3	51.5
	TAE(finetune) [20]	25.7	20.5	51.8	42.1	37.1	68.9	86.2	48.4	47.6
	Temporal Ranking [22]*	10.8	20.7	43.3	37.6	12.2	68.7	62.9	46.2	37.8
	Temporal Ranking(finetune) [22]	31.6	27.5	61.1	**53.4**	35.8	70.3	84.2	59.4	52.9
Ours	EmoCo	34.3	31.9	63.9	52.5	44.0	77.0	78.3	44.2	53.3
	EmoCo(finetune)	42.7	41.0	66.3	45.1	50.9	75.5	88.9	58.6	**58.6**

TABLE III

F1 SCORE (IN %) OF 12 AUS BY THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHODS ON THE BP4D DATASET. * MEANS THAT THE RESULTS ARE REPORTED IN THE ORIGINAL PAPERS.

	Methods/AU	1	2	4	6	7	10	12	14	15	17	23	24	Avg.
Supervised	DRML [37]*	**55.7**	**54.5**	**58.8**	56.6	61.0	53.6	60.8	57.0	**56.2**	50.0	53.9	53.9	56.0
	EAC [19]*	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
	JAA [29]*	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
	DSIN [6]*	51.7	40.4	56.6	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
	SRERL [16]*	46.9	45.3	55.6	77.1	**78.4**	83.5	87.6	63.9	52.2	**63.9**	**47.1**	53.3	**62.1**
Self-supervised	MoCo [11]	41.9	28.7	40.8	75.6	70.8	82.7	85.8	62.4	31	54.2	31.3	37.6	53.6
	MoCo(finetune) [11]	40.8	28.9	41	72.2	71	81.4	84.4	62	35.7	54.8	37.1	39.8	54.1
	Fab-Net [32]*	43.3	35.7	41.6	72.9	63.0	75.9	83.5	57.7	26.5	48.2	33.6	42.4	52.0
	Fab-Net(finetune) [32]	45.8	36.2	47.6	76.1	73.3	81.3	85.6	60.6	34.1	58.2	39.7	41.7	56.7
	TAE [20]*	47.0	45.9	50.9	74.7	72.0	82.4	85.6	62.3	48.1	62.3	45.9	46.3	60.3
	TAE(finetune) [20]	47.4	38	48.5	74.5	71.1	82.8	85.6	64	41.7	61.8	43.2	40.7	58.3
	Temporal Ranking [22]*	35.2	25.5	30.2	71.3	69.6	81.3	83.3	59.1	30.3	56.1	27.0	33.4	50.2
	Temporal Ranking(finetune) [22]	48.4	47	50.8	74.7	75.2	84.4	85.6	57.5	36.7	61.6	44.2	43.8	59.1
Ours	EmoCo	45.4	30.5	55.5	76.1	75.7	**84.4**	87.6	**66.6**	39.6	59.1	41.3	49.8	59.3
	EmoCo(finetune)	50.2	44.7	53.9	74.8	76.6	83.7	**87.9**	61.7	47.6	59.8	46.9	**54.6**	61.9

After finetuned, EmoCo achieves better performance than almost every supervised method, even the methods like JAA [29] and EAC [19] using facial landmarks as assistance to learn region-specific representations. Although SRERL [16] on BP4D dataset is an exception, it exceeds EmoCo by merely 0.2%, for the reason that SRERL [16] uses a larger backbone (VGG19 [30]) than EmoCo (ResNet-50 [12]) and uses a GCN to estimates the relationships among AUs. Apart from performance improvement, the fine-tuning of EmoCo is faster than directly training supervised methods. EmoCo converges in less than 5 epochs when being finetuned. The performance of EmoCo under linear protocol is even comparable to those supervised methods.

E. Ablation Study

We evaluate the effects of EmoCo's key components, including finetuning $\mathcal{F}$, the multi-queue dictionary MQ, contrastive loss $\mathcal{L}_{cont}$, emotion-classification loss $\mathcal{L}_{emo}$, and

TABLE IV
ABLATION STUDY OF THE KEY COMPONENTS OF EMOCO. $\mathcal{F}$: FINETUNING; MQ: MULTI-QUEUE DICTIONARY; $\mathcal{L}_{cont}$: CONTRASTIVE LOSS; $\mathcal{L}_{emo}$: EMOTION CLASSIFICATION LOSS; $\mathcal{N}$: INTRA-CLASS NORMALIZATION.

	$\mathcal{F}$	MQ	$\mathcal{L}_{cont}$	$\mathcal{L}_{emo}$	$\mathcal{N}$	$f1$
GFT [9]	–	✓	✓	✓	✓	53.0
	–	–	✓	✓	✓	48.6
	–	✓	✓	✓	–	47.5
	–	✓	✓	–	–	46.2
	–	✓	–	✓	–	47.6
	✓	✓	✓	✓	✓	**58.6**
	✓	–	✓	✓	✓	57.7
	✓	✓	✓	✓	–	57.3
	✓	✓	✓	–	–	53.3
	✓	✓	–	✓	–	52.3
DISFA [23]	–	✓	✓	✓	✓	53.3
	–	–	✓	✓	✓	50.2
	–	✓	✓	✓	–	48.4
	–	✓	✓	–	–	41.4
	–	✓	–	✓	–	47.6
	✓	✓	✓	✓	✓	**58.6**
	✓	–	✓	✓	✓	57.2
	✓	✓	✓	✓	–	57.0
	✓	✓	✓	–	–	54.0
	✓	✓	–	✓	–	53.6
BP4D [34]	–	✓	✓	✓	✓	59.3
	–	–	✓	✓	✓	57.8
	–	✓	✓	✓	–	56.2
	–	✓	✓	–	–	55.0
	–	✓	–	✓	–	58.5
	✓	✓	✓	✓	✓	**61.9**
	✓	–	✓	✓	✓	60.6
	✓	✓	✓	✓	–	60.7
	✓	✓	✓	–	–	59.6
	✓	✓	–	✓	–	58.9

intra-class normalization $\mathcal{N}$ [1]. $\mathcal{F}$ stands for whether to finetune pretrained EmoCo on target datasets. MQ means each queue is for per class, otherwise one queue is for all classes. $\mathcal{L}_{cont}$ stands for whether to use emotion-aware contrastive loss when pretraining. $\mathcal{L}_{emo}$ stands for whether to use coarse emotion supervision. And $\mathcal{N}$ denotes using intra-class normalization [1] or L_2 normaliztion. Table IV reports the F1 score of EmoCo on three datasets when applying different components. We conclude the observations below.

Firstly, EmoCo with intra-class normalization [1] effectively combines coarse emotion supervision and emotion-aware contrastive learning to learn fine-grained AU representations. When utilizing $\mathcal{L}_{emo}$ alone, EmoCo reduces to a plain emotion classfication network. Similarly, when applying $\mathcal{L}_{cont}$ alone, EmoCo reduces to multi-queue MoCo [11]. Each of them either only focuses on classifying emotional categories or forces the model to learn differences in the background or something emotion-unrelated, resulting in the peformance reduction. Besides, without intra-class normalization, directly combining $\mathcal{L}_{emo}$ and $\mathcal{L}_{cont}$ sometimes performs even worse than only using $\mathcal{L}_{emo}$, proving the effectness of intra-class normalization that relieves the conflict between $\mathcal{L}_{emo}$ and $\mathcal{L}_{cont}$ from the negative side.

Secondly, choosing negative samples from the same emotion class of the positive samples is more helpful to EmoCo than randomly sample selection. In this way, EmoCo introduces hard negative samples and encourages contrastive learning to find more fine-grained emotional differences between images.

Thirdly, after finetuned on target dataset, the performance of EmoCo can be dramatically improved.

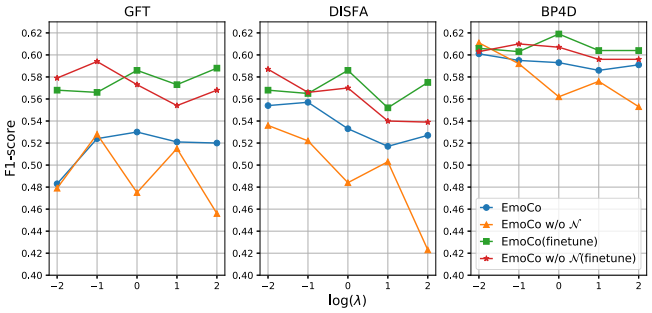


Fig. 4. F1-scores of EmoCo with different λs which balance emotion-classification loss $\mathcal{L}_{emo}$ and emotion-aware contrastive loss $\mathcal{L}_{cont}$. $\mathcal{N}$ stands for intra-class normalization. With $\mathcal{N}$, the performance of EmoCo is stable when λ varies.

F. Balance of $\mathcal{L}_{emo}$ and $\mathcal{L}_{cont}$

When training EmoCo, we adopt a weight coefficient λ to balance the emotion-classification loss $\mathcal{L}_{emo}$ and the emotion-aware contrastive loss $\mathcal{L}_{cont}$. In this section, we explore the effects of different λ on the performance of EmoCo. The results are showed in Fig. 4. By analyzing the experimental results, we find that: (a) The value of λ has little influence on the performance of EmoCo under finetuning protocol, no matter EmoCo has an intra-class normalization [1] module or not. (b) The performance of EmoCo with intra-class normalization [1] under linear protocol is more stable than the one with L_2 normalization. (c) EmoCo with intra-class normalization [1] achieves superior performance without special search on λ. We attribute this to the fact that intra-class normalization [1] could integrate the prototype of each emotional category $\tilde{\mathbf{w}}$ into the contrastive loss, which adaptively adjusts the influence between $\mathcal{L}_{emo}$ and $\mathcal{L}_{cont}$.

G. Visualization

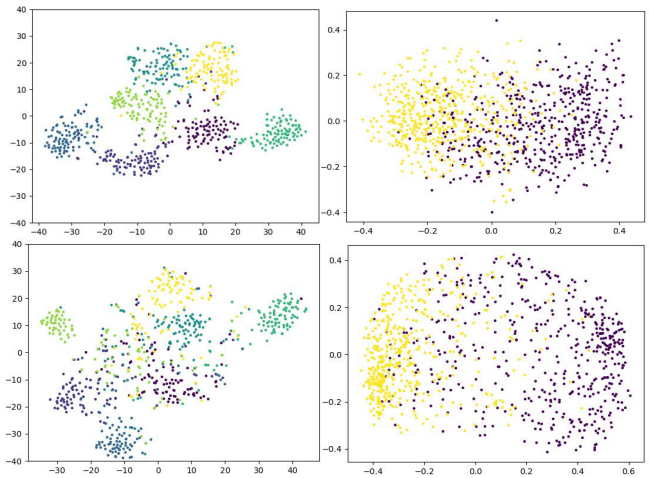


Fig. 5. Visualization of the learned features. **Top**: features with emotion-supervision only; **Bottom**: features of EmoCo; **Left**: Colors indicate 7 emotional categories; **Right**: Colors indicates whether AU6 exists.

In this section, we visualize the learned features of EmoCo and the emotion-supervised baseline. The left half of Fig. 5 shows the feature distribution of the seven emotional categories. The features of both models are separated into seven

classes, while the emotion class boundaries of EmoCo's features are less clear. The right half of Fig. 5 shows the intra-class feature distribution within an emotion class. The features of EmoCo (lower) are more separable than the features of expression-supervised baseline. Fig. 5 demonstrates that the EmoCo-learned feature are more AU discriminative than the features learned with emotion only.

V. CONCLUSIONS

In this work, we propose an Emotion-aware Contrastive Learning framework to learn discriminative AU representations leveraging large-scale expression dataset. Regarding AUs as finer descriptors of facial behaviors than emotional categories, EmoCo adopts a coarse-to-fine contrastive learning framework [1] to separate emotional categories and distinguish instance-level features simultaneously. EmoCo achieves comparable performance to the state-of-the-art AU detection methods under both linear and finetuning protocol. We will modify this framework into a self-supervised version combining deep cluster [2] and a novel normalization method in the future.

REFERENCES

[1] G. Bukchin, E. Schwartz, K. Saenko, O. Shahar, R. Feris, R. Giryes, and L. Karlinsky. Fine-grained angular contrastive learning with coarse labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8730–8740, 2021.

[2] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018.

[3] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[6] C. Corneanu, M. Madadi, and S. Escalera. Deep structure inference network for facial action unit recognition. In *ECCV*, pages 298–313, 2018.

[7] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[8] P. Ekman and W. V. Friesen. Manual of the facial action coding system (facs). *Trans. ed. Vol. Consulting Psychologists Press, Palo Alto*, 1978.

[9] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 581–588. IEEE, 2017.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[15] M. Lewis, J. M. Haviland-Jones, and L. F. Barrett. *Handbook of emotions*. Guilford Press, 2010.

[16] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019.

[17] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2584–2593. IEEE, 2017.

[18] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusingshao. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2017.

[19] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 103–110. IEEE, 2017.

[20] Y. Li, J. Zeng, and S. Shan. Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[21] Y. Li, J. Zeng, S. Shan, and X. Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019.

[22] L. Lu, L. Tavabi, and M. Soleymani. Self-supervised learning for facial action unit recognition through temporal consistency. In *BMVC*, 2020.

[23] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[24] A. Mollahosseini, B. Hasani, and M. 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.

[25] X. Niu, H. Han, S. Shan, and X. Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. *arXiv preprint arXiv:1910.11012*, 2019.

[26] G. Peng and S. Wang. Weakly supervised facial action unit recognition through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2188–2196, 2018.

[27] G. Pons and D. Masip. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*, 2018.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[29] Z. Shao, Z. Liu, J. Cai, and L. Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *ECCV*, pages 705–720, 2018.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] C. Wang, J. Zeng, S. Shan, and X. Chen. Multi-task learning of emotion recognition and facial action unit detection with adaptively weights sharing network. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 56–60. IEEE, 2019.

[32] O. Wiles, A. Koepke, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*, 2018.

[33] S. Wu, S. Wang, B. Pan, and Q. Ji. Deep facial action unit recognition from partially labeled data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3951–3959, 2017.

[34] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition*, pages 1–6. IEEE, 2013.

[35] Y. Zhang, W. Dong, B.-G. Hu, and Q. Ji. Classifier learning with prior probabilities for facial action unit recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5108–5116, 2018.

[36] K. Zhao, W.-S. Chu, and A. M. Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 2090–2099, 2018.

[37] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *CVPR*, pages 3391–3399, 2016.