Knowledge-Driven Self-Supervised Representation Learning for Facial Action Unit Recognition

Yanan Chang, Shangfei Wang*
University of Science and Technology of China, Hefei, Anhui, China
cyn123@mail.ustc.edu.cn, sfwang@ustc.edu.cn

Abstract

Facial action unit (AU) recognition is formulated as a supervised learning problem by recent works. However, the complex labeling process makes it challenging to provide AU annotations for large amounts of facial images. To remedy this, we utilize AU labeling rules defined by the Facial Action Coding System (FACS) to design a novel knowledge-driven self-supervised representation learning framework for AU recognition. The representation encoder is trained using large amounts of facial images without AU annotations. AU labeling rules are summarized from FACS to design facial partition manners and determine correlations between facial regions. The method utilizes a backbone network to extract local facial area representations and a project head to map the representations into a low-dimensional latent space. In the latent space, a contrastive learning component leverages the inter-area difference to learn AU-related local representations while maintaining intra-area instance discrimination. Correlations between facial regions summarized from AU labeling rules are also explored to further learn representations using a predicting learning component. Evaluation on two benchmark databases demonstrates that the learned representation is powerful and data-efficient for AU recognition.

1. Introduction

Facial AUs defined by the Facial Action Coding System [4] describe the activities of sets of specific facial muscles. Nearly all facial behaviors can be represented through one or more AUs. Automatic facial AU recognition has attracted attention due to its potential in a wide variety of applications.

The majority of current works on facial AU recognition are supervised, requiring fully AU-labeled images for training. In general, there are two different approaches to supervise AU recognition. The first treats AU recognition as a multi-label classification problem to be solved by directly constructing an end-to-end deep network [2, 5]. However, AUs are typically correlated to partial facial areas. These works only utilize global facial information for AU recognition, limiting their performance. Recent works [14, 15, 17, 30] have opted for the second approach, which tries to learn more AU-specific patterns to enhance AU recognition. For example, these works locate AUs based on facial landmarks and muscles. The nearby areas of specific AUs are used to predict their labels. Leveraging more AU-specific patterns can effectively improve AU recognition. However, AU labels must be annotated by experienced experts, which is time and labor intensive. Existing AU-labeled databases are too limited to take advantage of these supervised methods.

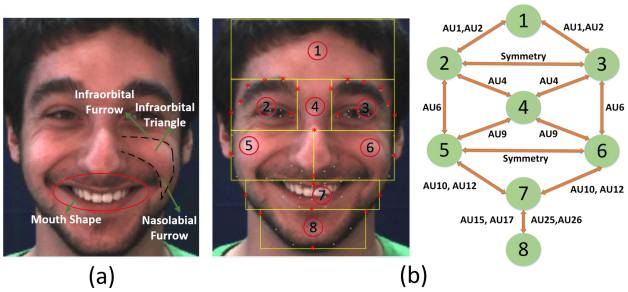


Figure 1. (a)The judgment areas of AU12 are shown. (b)Left: Facial areas are divided into eight parts according to AU-related appearance changes. Right: The relationships between facial areas. Corresponding AUs are labeled on the edges.

Recently, several works have tackled the issue of AU annotations. Some works [25, 27] try to perform semi-supervised AU recognition. These works summarize label distribution from ground-truth AU labels, and use the learned distribution to improve AU recognition. However, the summarized patterns may not be consistent with the true distribution due to limited ground-truth AU labels. Li *et al.* [16] have created self-supervised learning methods in which large amounts of unlabeled images are used to learn representations. They utilize the transformation between two adjacent frames as the supervisory signal to learn AU-related

*This is the corresponding author.

global facial representations, ignoring the local property of AUs. There are also several self-supervised methods [1, 6] that learn powerful visual representations for image classification via contrastive learning. However, both Li *et al*. [16] and the contrastive learning works design self-supervised tasks using random augmentation or temporal information. They do not fully leverage task-related domain knowledge.

To address these obstacles, we propose a novel knowledge-driven self-supervised representation learning framework for AU recognition to alleviate the demand for AU labels. Specifically, we first summarize AU labeling rules taken from FACS as domain knowledge. FACS determines AUs according to different facial appearance changes. For example, as shown in Figure 1a, some key facial appearance changes of AU12 consist of lip corners raising, infraorbital triangle raising, and so on. AU-related appearance changes are summarized, and facial areas are divided into eight sections according to the locations of the appearance changes, as shown in the left image of Figure 1b. There are also correlations between local areas, summarized in the right image of Figure 1b. The summarized knowledge is leveraged to design a self-supervised representation learning framework. A backbone network extracts local representations for each facial part, and a project head maps the local features into a low-dimensional latent space. In the latent space, a contrastive learning component and a predicting learning component train the feature encoder. The challenge with contrastive learning is designing reasonable data pairs. Positive pairs pull closer and negative pairs push apart. In our contrastive learning component, the embeddings from the same and symmetrical areas are treated as positive pairs according to AU labeling rules. All others are regarded as negative pairs. In addition, for each area, the embeddings from the same input image are treated as positive pairs to maintain intra-area instance discrimination. We propose a predicting learning component to leverage the summarized inter-area relationships to enhance representation learning. A group of predictors are used to learn correlations between the embeddings from different areas in the latent space. The inter-area correlations are utilized as supervisory signals. Finally, the proposed representation learning framework is trained on a large available unlabeled database. AU classifiers are further trained on two benchmark databases to evaluate the efficacy of the learned representations for AU recognition.

The contributions of the paper can be summarized as follows. We propose a novel knowledge-driven self-supervised representation learning framework for AU recognition, which can learn AU-related local representations from large amounts of available unlabeled images. Unlike previous self-supervised learning methods ignoring task-related domain knowledge, we leverage both the difference and correlation between local facial areas as supervisory signals under the guidance of AU labeling rules. Evaluation on two benchmark databases shows that the learned local features are powerful and data-efficient for AU recognition compared to state-of-the-art self-supervised, semi-supervised, and supervised methods.

2. Related Work

2.1. AU recognition

A comprehensive survey on facial AU recognition can be found in [18]. In this section, we provide a brief review of advances in facial AU recognition.

The majority of recent works on AU recognition are based on supervised methods. Several works [2, 5] treat AU recognition as a multi-label recognition problem, and have achieved better performance than hand-crafted feature extractors. However, AUs are usually only related to partial facial areas. These works ignore the property, which has limited their performance. Recent works prefer to enhance AU recognition by learning AU-specific patterns. For example, Zhao *et al*. [30] try to address region and multi-label learning jointly. Li *et al*. [15] combine region of interest (ROI) adaptation with optimal LSTM-based temporal fusing. Shao *et al*. [22] use an adaptive attention module to extract precise local features in their joint AU detection and face alignment framework (JÂA-Net). Li *et al*. [14] integrate semantic correlations between AUs to create a deep region learning framework for AU recognition. Jacob *et al*. [11] leverage transformer encoder to perform AU recognition. AU-specific representations are extracted based on the regions of interest. Tang *et al*. [26] learn pixel-level attention to enhance AU recognition by a pixel-interested learning method. Song *et al*. [23, 24] leverage graph neural network to exploit AU correlations to enhance AU recognition. These works benefit from leveraging more AU-related patterns. However, supervised works need fully AU-labeled data for training. A lack of available AU-labeled databases has limited their generalization.

Some works try to alleviate the demand for AU annotations. Several perform semi-supervised AU recognition by summarizing label distributions from ground-truth AU labels. For example, Song *et al*. [25] marginalize over the latent values to tackle missing labels during inference for their Bayesian Group-Sparse Compressed Sensing (BGCS) method. Wu *et al*. [27] introduce a semi-supervised AU recognition method (DAU-R) in which AU distributions are captured by restricted Boltzmann machine (RBM). Niu *et al*. [20] propose semi-supervised AU recognition method by leveraging unlabeled web face images. However, the label distribution learned from the limited ground-truth AU labels may not be consistent with their true distribution. Li *et al*. [16] propose self-supervised representation learning for AU recognition. The transformation between two facial

Table 1. The appearance changes and judgment areas for AUs.

AU	AU Name	Appearance Changes	Judgment Areas
1	Inner Brow Raiser	There are wrinkles in the middle of the forehead; Inner eyebrows raise.	1, 2, 3
2	Outer Brow Raiser	Lateral eyebrows raise; Lateral eye cover folds are stretched upwards; There are wrinkles in the lateral portion of forehead.	1, 2, 3
4	Brow Lowerer	Eyebrows are lowered and pulled closer together; There are wrinkles or muscle bulges between eyebrows.	2, 3, 4
6	Cheek Raiser	There are crow's feet lines; Infraorbital triangles are lifted upwards.	2, 3, 5, 6
7	Lid Tightener	Eye apertures are narrowed; Lower eyelids are stretched and raised.	2, 3
9	Nose Wrinkler	Infraorbital triangles are lifted upwards; There are wrinkles across nose root.	4, 5, 6
10	Upper Lip Raiser	There exist pouches in inner corner of infraorbital triangle; Center of upper lip is lifted up.	5, 6, 7
12	Lip Corner Puller	Lip corners are pulled obliquely; Nasolabial furrows are deepened; Infraorbital triangles are lifted upwards and infraorbital furrows are deepened.	5, 6, 7
14	Dimpler	Lip corners are tightened; There are wrinkles and bulges at mouth corners.	7
15	Lip Corner Depressor	Lip corners are moved down; Lip shape is stretched down; There exist bulges below the lip corners.	7, 8
17	Chin Raiser	The lower lip are lifted up; And there are wrinkles on the chin boss.	7, 8
23	Lip Tightener	Lips are narrowed and tightened.	7
24	Lip Presser	There are evidences of lips pressing together.	7
25	Lips Part	There are evidences of lips parting and teeth exposing.	7, 8
26	Jaw Drop	There are appearances of mandible lowering by relaxation; There is space between the upper and lower teeth.	7, 8

images is used as the supervisory signal to learn AU-related representations. Large amounts of unlabeled videos can be used to train the framework. However, Li *et al*. [16]'s work only learns global facial features and ignores the local property of AUs. Task-related domain knowledge is not fully explored, and these factors have limited its performance.

In this paper, we propose a knowledge-driven self-supervised representation learning framework for AU recognition that can learn powerful AU-related local representations from unlabeled facial images.

2.2. Self-supervised learning

Great progress has recently been made in self-supervised learning, which adopts supervisory signals of the data itself to learn representations from large amounts of unlabeled data. The most competitive self-supervised representation learning method is contrastive learning [1,6,9,12,21], which utilizes contrastive loss to force low-dimensional data embeddings to pull together positive data pairs and push apart negative data pairs. The key question of contrastive learning is how to design reasonable data pairs.

Hénaff *et al*. [9] leverage contrastive predictive coding by dividing images into overlapping patches. He *et al*. [6] and Chen *et al*. [1] perform contrastive self-supervised representation learning by generating different image views under random data augmentation. Khosla *et al*. [12] propose supervised contrastive learning, which considers the label information. These works typically leverage the difference between data pairs using random augmentation or patch division. However, they do not fully utilize task-related domain knowledge.

In this paper, we utilize domain knowledge to guide the design of a self-supervised learning framework. For the contrastive learning component, data pairs are designed according to AU labeling rules, leveraging the inter-area difference to supervise representation learning. The correspondences between facial areas are used as supervisory signals to further enhance the learned representation via a predicting learning component.

3. Problem Statement

Let $D = \{x_i\}_{i=1}^{T}$ denote training samples without AU annotations, where x_i represents a facial image, and T is the number of all the training samples. The goal is to learn a function $f(\cdot)$ from D to extract AU-related local representations for each input image.

4. Methodology

AU labeling rules are summarized from the FACS. That knowledge is leveraged to design a self-supervised representation learning framework for AU recognition.

4.1. AU labeling rules

AUs describe facial muscular activities. The FACS details how to recognize AUs and label AU intensities via facial appearance changes. For example, AU1 represents inner brow raiser. Important appearance changes for AU1 include the raising of the inner brow and the appearance of wrinkles in the center of the forehead. AU activation triggers appearance changes in different facial areas. Table 1 presents the main appearance changes for 15 common AUs.

We split the global facial regions into eight separate judgment areas according to the emerging locations of appearance changes and facial landmarks, as shown in the left side of Figure 1b. Each facial area is a rectangular box, located by several special landmarks labeled with a red asterisk. Table 1 summarizes the judgment areas for each AU. For example, the judgment areas related to AU1 include areas 1, 2, and 3. Each AU can be judged by jointly observing the appearance changes of several different judgment areas. Facial areas will change appearance as different AUs are activated.

There are also correlations between different facial areas.

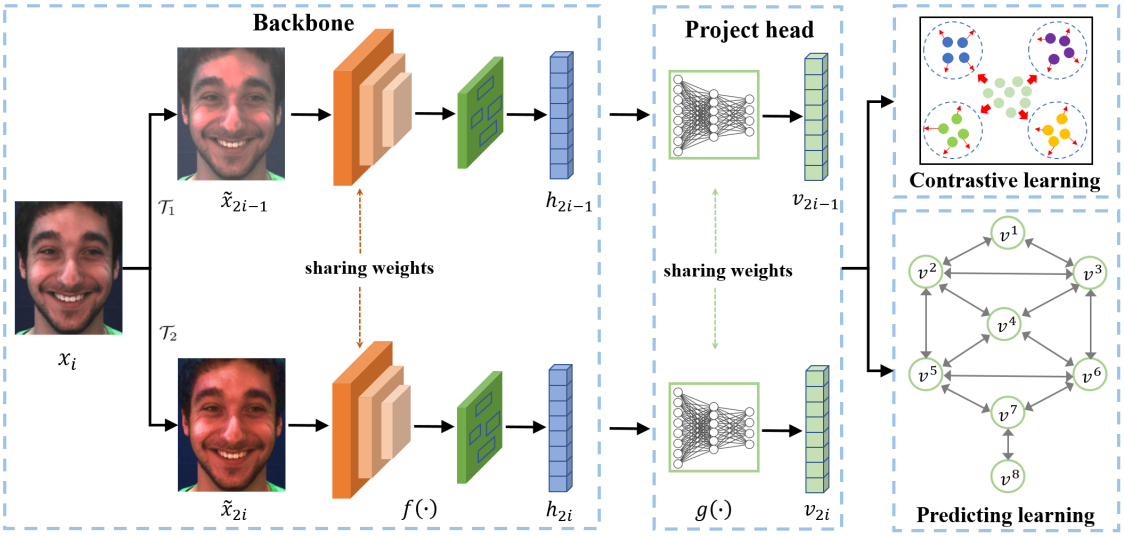


Figure 2. The framework. In the backbone network, one input facial image is first transformed into two different views by different augmentations. Then $f(\cdot)$ based on CNNs and a ROI Align layer is utilized to extract eight local representations for each view. Project head maps the local facial features to a low-dimensional latent space by $g(\cdot)$. In the latent space, contrastive learning and predicting learning components are introduced to train the framework in a self-supervised manner.

On the one hand, appearance changes between asymmetric areas correspond. For example, both AU12 and AU10 will cause appearance changes in judgment areas 5, 6, and 7. The medial portion of upper lip raising in area 7 and the upper portion of nasolabial furrow deepening in areas 5 and 6 will occur simultaneously while AU10 is activated. Lip elongating and angling obliquely at the corner in area 7 will occur simultaneously with the lower portion of nasolabial furrow deepening in areas 5 and 6 due to the activation of AU12. This demonstrates that the appearance changes in areas 5, 6, and 7 are highly corresponding owing to the activation of different AUs. On the other hand, according to [3], facial appearance changes on the left and right sides are usually similar in deliberate and emotional facial actions due to symmetry determined by facial muscular mechanisms, though their intensities may be different. For example, while AU1 is activated, the appearance changes presented in areas 2 and 3 are usually similar, though the magnitude of inner brow raising may be different. The inter-area relation graph is summarized in the right image of Figure 1b, in which each vertex denotes one judgment area. Related AUs are marked on the edges connecting related vertices.

4.2. The proposed representation learning framework

Figure 2 shows the representation learning framework, including a backbone network, a project head, a contrastive learning component, and a predicting learning component. First, the input facial image is transformed into two views by different augmentation methods. Then the two transformed views are respectively fed into $f(\cdot)$ to extract local representations for the separate facial regions defined in the left image of Figure 1b. In order to leverage AU labeling rules to train the feature encoder in a self-supervised manner, a project head maps the local facial representations to a low-dimensional latent space, as in other works [1, 12]. Finally, under the guidance of summarized knowledge, a contrastive learning component and a predicting learning component are designed in the latent space to train the representation learning framework.

4.2.1 Contrastive learning component

The facial area can be divided into eight parts according to the summarized labeling rules. Representations differ from one area to the next. The contrastive learning component differentiates the representations from the facial areas according to the summarized knowledge. First, a mini-batch of N images $\{x_i\}_{i=1}^{N}$ is randomly sampled from D. The related mini-batch for training includes 2N samples, $\{\tilde{x}_i\}_{i=1}^{2N}$. $\tilde{x}_{2i}$ and $\tilde{x}_{2i-1}$ are augmentations of x_i (i=1...N). After passing through $f(\cdot)$, $16N$ local representations are acquired for the mini-batch, $\{h_i^1, h_i^2, ..., h_i^8\}_{i=1,...,2N}$. Then a project head $g(\cdot)$ is applied to map the local representations into a low-dimensional latent space, $\{v_i^1, v_i^2, ..., v_i^8\}_{i=1,...,2N}$. The embeddings from the same or symmetrical judgment regions are treated as positive pairs in the latent space; otherwise, they are regarded as negative pairs. The cosine similarity between v and u can be denoted as $csim(v,u) = \frac{v^T u}{\|v\|_2 \|u\|_2}$. We introduce the following loss functions:

$$L^a = \sum_{i=1}^{2N} \sum_{p=1}^{8} \ell_{i,p}^a \tag{1}$$

$$\ell_{i,p}^a = \frac{1}{N_{i,p}} \sum_{j=1}^{2N} \sum_{q=1}^{8} \mathbb{1}_{[i \neq j \vee p \neq q]} \cdot \mathbb{1}_{[q \in \Phi(p)]} \cdot \ell(v_i^p, v_j^q) \tag{2}$$

$$\ell(v_i^p, v_j^q) = -\log \frac{\exp\left(\text{csim}\left(\boldsymbol{v}_i^p, \boldsymbol{v}_j^q\right)/\tau\right)}{\sum_{k=1}^{2N} \sum_{r=1}^{8} \mathbb{1}_{[i \neq k \vee p \neq r]} \cdot \exp\left(\text{csim}\left(\boldsymbol{v}_i^p, \boldsymbol{v}_k^r\right)/\tau\right)} \tag{3}$$

where $N_{i,p}$ is the total number of positive pairs related to the embedding v_i^p. $\mathbb{1}_{[\cdot]} \in \{0, 1\}$ is a function evaluating to 1 if the condition is true. $\Phi(p)$ denotes the set of areas, including p and its symmetric area (if the symmetric area exists). τ is a temperature parameter. The above loss function pulls positive data pairs closer and pushes negative data pairs apart. However, it tends to reduce the diversity of intra-area representations. Intra-area contrastive learning is used to counteract this. For each local facial area, the embeddings from the same input facial image are treated as positive pairs; otherwise, they are regarded as negative pairs. The loss functions are defined as follows:

$$L^b = \sum_{m=1}^{8} \ell_m^b \tag{4}$$

$$\ell_m^b = -\sum_{i=1}^{2N} \log \frac{\exp\left(\text{csim}\left(\boldsymbol{v}_i^m, \boldsymbol{v}_{\psi(i)}^m\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \cdot \exp\left(\text{csim}\left(\boldsymbol{v}_i^m, \boldsymbol{v}_k^m\right)/\tau\right)} \tag{5}$$

where $\psi(i)$ is the index of the other augmented image from the same input facial image. The overall contrastive learning loss is defined as Equation-6, where λ is a weight to balance the degree of the intra-area instance diversity. Overall, the contrastive learning component leverages the inter-area differences as supervisory signals, while keeping the intra-area representations diverse.

$$L_{con} = L^a + \lambda L^b \tag{6}$$

4.2.2 Predicting learning component

There are correspondences between local facial regions due to the activation of AUs. The representation of one judgment area should be predictive from its related areas due to the inter-area co-occurrence relations of the appearance changes. In this section, we introduce a predicting learning component that leverages these relationships to enhance representation learning. A group of predictors is applied in the latent space to learn the relationships between the related facial areas. Figure 2 shows the predicting graph between local embeddings. The group of predictors is denoted as $\{\varphi_{qp}\}$, where φ_{qp} denotes a predictor predicting embeddings of area p from the ones of related area q. There is one predictor for each arrow. In total, 26 predictors are adopted to learn the correlations between local facial areas. The predicting relations are shown in Equation-7, where v_i^q denotes the q-th latent feature of the i-th image $\tilde{x}_i$ in the mini-batch. K denotes the number of the related areas of area p. $\mathbb{1}_{[q \sim p]}$ is a function evaluating to 1 if p relates to q. $\hat{v}_i^p$ denotes the predicted embeddings of p-th areas of i-th image $\tilde{x}_i$ in the mini-batch. For area p, the multiple predicted representations from other related areas of the same image are averaged as the final predicted results.

$$\hat{v}_i^p = \frac{1}{K} \sum_{q=1}^{8} \mathbb{1}_{[q \sim p]} \cdot \varphi_{qp}(v_i^q) \tag{7}$$

$$L_{pre} = \sum_{i=1}^{2N} \sum_{p=1}^{8} \left(1 - \frac{v_i^p \cdot \hat{v}_i^p}{\|v_i^p\|_2 \cdot \|\hat{v}_i^p\|_2}\right) \tag{8}$$

The distance between the predicted and target embeddings is closed by one cosine loss. The function is shown as Equation-8. The loss function forces the predicted embeddings to be close to the target one. The correlations between representations from different areas are exploited to supervise the training of the representation learning framework.

4.2.3 Overall learning

The overall learning loss is defined as Equation-9, where α and β are weighted coefficients to balance the contrastive and predicting components.

$$L = \alpha L_{con} + \beta L_{pre} \tag{9}$$

The contrastive learning component tries to leverage the inter-area difference to train the feature encoder while maintaining intra-area instance diversity. The predicting learning component leverages the correlations between different areas to further enhance the learned representations. By jointly training the contrastive and predicting learning components, our method can differentiate the representations of different areas while learning the inter-area co-occurrence relations. Both the difference and correspondence between facial areas are leveraged as supervised signals to improve learning of AU-related local representations. AU labeling rules guide the design of the self-supervised representation learning framework.

5. Experiments

5.1. Experimental conditions

The framework is trained on the BP4D+ database [29]. After training, learned representations are evaluated by training AU classifiers on two databases: the BP4D database [28] and the Denver Intensity of Spontaneous Facial Action database (DISFA) [19].

The BP4D+ database is a multi-modal spontaneous emotion corpus consisting of 140 subjects. Each participant performs with 10 tasks. There are about 1.4 million frames in total. All available 2D image samples are used. The subjects are randomly divided, with 70% used for training and 30% used for validating.

The BP4D database includes spontaneous facial videos of eight tasks from 41 subjects. These subjects are different from those recorded for the BP4D+ database. There are 328 two-dimensional videos coded with 12 AUs (i.e., AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, and AU24). We use all available AU labels and around 140,000 valid image samples.

The DISFA database consists of spontaneous videos of 27 subjects. The AUs are labeled with intensities ranging from 0 to 5. There are about 130,000 valid image samples. Eight AUs (i.e., AU1, AU2, AU4, AU6, AU9, AU12, AU25, and AU26) are considered and each AU with an intensity greater than or equal to 2 is treated as active.

Dlib toolkit [13] is applied to detect 68 facial landmarks for each image. The augmentation methods are random color distortions with different brightness, saturation, constrast, and hue. $f(\cdot)$ includes CNNs and an RoI-align layer [7]. The CNNs are based on ResNet-50 network [8]. The features extracted from the final convolutional layer of the conv4_x are utilized as global features, as is common practice in other works [7, 10]. $f(\cdot)$ takes a 224 x 224 RGB image as the input and outputs a 4096-dimensional local representation for each facial area. $g(\cdot)$ is a multi-layer perceptron (MLP) with a hidden layer of size 2048; it outputs vectors of size 128. In the latent space, a group of predictors is applied to learn the vector correspondences. Each predictor is an MLP with a single hidden layer of size 1024. The representation learning framework is trained end-to-end by minimizing the loss functions in Equation-9. λ, τ, α, and β are set to 0.1, 0.07, 0.01, and 1, respectively. The framework is implemented by PyTorch and trained by Adam optimizer with an initial learning rate of 0.0001 and batch size of 128.

In order to evaluate the learned representation, the AUs are divided into several groups according to their related judgment areas from Table 1. AUs are usually judged by jointly observing several related judgment areas. Table 2 shows the groups of predicted AUs and their related facial judgment areas. AU classifiers based on MLP with two hidden layers of sizes 1024 and 128 are trained for each group via cross-entropy loss (fixed the parameters of $f(\cdot)$). Subject-independent 3-fold cross-validation is used for the BP4D and DISFA databases. F1 score is adopted to evaluate the performance of AU recognition.

Table 2. The correspondence between facial areas and predicted AUs.

Facial areas	Predicted AUs
1, 2, 3, 4	AU1, AU2, AU4
2, 3, 5, 6	AU6, AU7
4, 5, 6, 7	AU9, AU10, AU12
7	AU14, AU23, AU24
7, 8	AU15, AU17, AU25, AU26

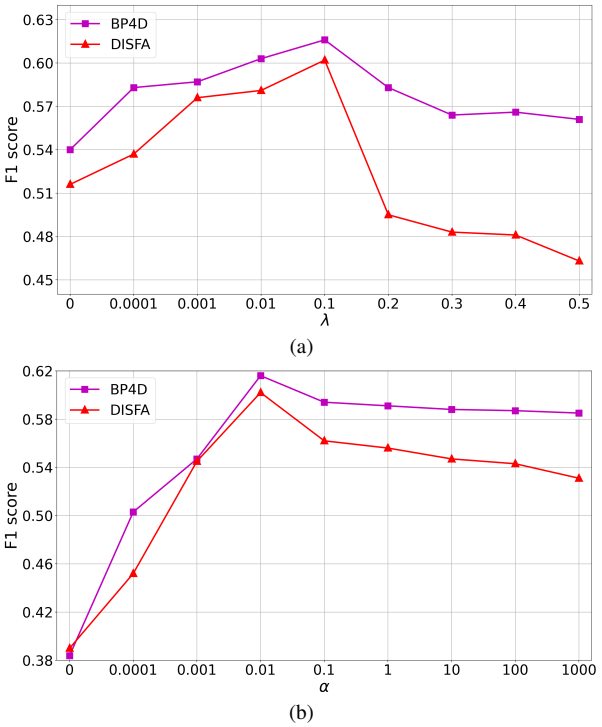


Figure 3. (a) F1 score with different λ on the BP4D and DISFA databases. (b) F1 score with different α on the BP4D and DISFA databases.

5.2. Experimental results and analysis

5.2.1 Evaluation for contrastive learning component

As shown in Equation 6, the inter-area contrastive learning loss L^a tries to differentiate the embeddings from different facial areas, but tends to reduce the diversity of intra-area representations. L^b tries to retain intra-area instance diversity, but also reduces the inter-area difference to a certain extent. λ is a weight to balance the degree of the intra-area instance diversity. When λ is 0, the intra-area loss L^b is discarded and no intra-area instance diversity is considered. As λ increases, the weight of L^b increases. We train the representation learning framework with different λ on

Table 3. F1 score of self-supervised AU recognition on the BP4D database.

Methods	SimCLR	MoCo	TCAE	**Ours**
AU1	11.3	7.6	43.1	50.1
AU2	6.0	2.9	32.2	45.4
AU4	19.7	13.7	44.4	53.6
AU6	67.5	79.2	75.1	79.2
AU7	72.2	79.8	70.5	78.4
AU10	81.1	85.0	80.8	85.2
AU12	75.8	87.7	85.5	87.4
AU14	52.6	61.6	61.8	65.4
AU15	16.0	33.7	34.7	51.5
AU17	22.3	56.8	58.5	56.1
AU23	4.5	16.3	37.2	44.6
AU24	9.4	28.8	48.7	42.0
Avg.	36.5	46.1	56.1	61.6

the BP4D+ database and evaluate the learned representations by training AU classifiers on the BP4D and DISFA databases. F1 scores on two databases with different λ are shown in Figure 3a. When λ is 0.1, which is the optimal result, the F1 score increases by 7.6% and 8.6% on the BP4D and DISFA databases, respectively, compared to when λ is 0. The results demonstrate that leveraging the inter-area difference while maintaining intra-area instance discrimination is effective for the representation learning. When λ increases beyond 0.1, F1 scores tend to decrease. The contrastive learning component is proficient at balancing the differences of inter-area embeddings and the diversity of intra-area embeddings.

5.2.2 Evaluation for the complementary of the contrastive and predicting learning components

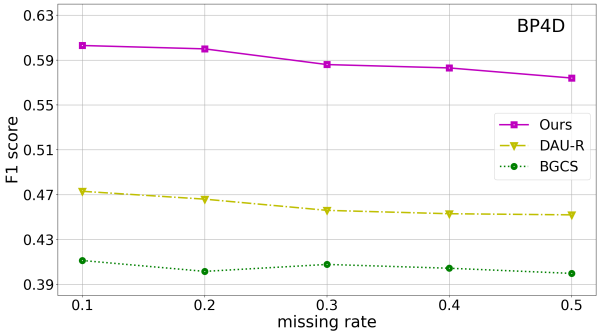
In this section, we evaluate the contrastive and predicting learning components by setting different α in Equation 9, while β is equal to 1. When α is 0, the contrastive learning component is discarded. Under this setting, the feature encoder is only trained by the predicting learning component. As α increases, the weight of the contrastive learning component increases. Figure 3b shows the evaluation results on the BP4D and DISFA databases. From Figure 3b, when α is 0.01, the performance increases significantly compared to when α is 0. It demonstrates that leveraging inter-area difference is important. When α increases beyond 0.01, the results tend to decrease due to the smaller weight of predicting learning component. This result demonstrates that the contrastive learning and predicting learning components can complement to each other. Balancing two components benefits the representation learning.

5.3. Comparison with self-supervised methods

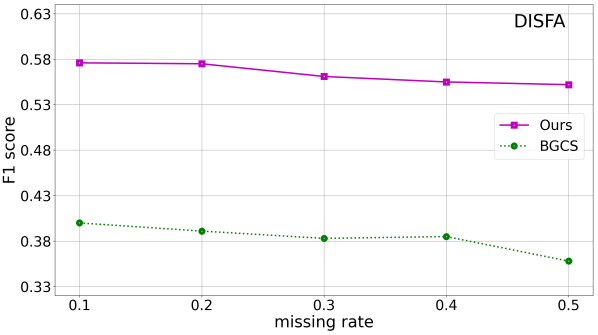
We compare our method with several self-supervised methods, including SimCLR [1], MoCo [6], and TCAE [16]. SimCLR and MoCo models are retrained on the BP4D+ database. Both models are based on ResNet-50 architecture. TCAE provides evaluation on the BP4D and DISFA databases. Their experimental results are used as a direct comparison. Tables 3 and 4 show the self-supervised results on the BP4D and DISFA databases.

Table 4. F1 score of self-supervised AU recognition on the DISFA database.

Methods	SimCLR	MoCo	TCAE	**Ours**
AU1	23.8	8.9	15.1	54.6
AU2	20.3	16.5	15.2	53.6
AU4	42.9	55.9	50.5	58.1
AU6	35.1	48.4	48.7	52.5
AU9	16.4	20.2	23.3	45.5
AU12	61.3	72.1	72.1	77.6
AU25	70.3	84.9	82.1	86.9
AU26	32.3	13.5	52.9	53.2
Avg.	37.8	40.0	45.0	60.2



(a)



(b)

Figure 4. (a) F1 score of semi-supervised AU recognition on the BP4D database. (b) F1 score of semi-supervised AU recognition on the DISFA database.

The proposed method achieves 25.1% and 22.4% improvement over SimCLR on the BP4D and DISFA databases, respectively. Our method is also 15.5% and 20.2% better than MoCo on those databases. Compared to the two contrastive learning methods, the proposed method makes full use of domain knowledge to guide the design of self-supervised tasks, improving task-related representation learning. Our method is also superior to TCAE, achieving F1 scores that are 5.5% and 15.2% higher on the BP4D and DISFA databases, respectively. TCAE ignores the local property and domain knowledge of AUs. Our method leverages AU labeling rules to design a self-supervised frame-

Table 5. F1 score of supervised AU recognition on the BP4D database.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	**Avg.**
DRML	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
ROI	36.2	31.6	43.4	77.1	73.7	85.0	87.0	62.6	45.7	58.0	38.3	37.4	56.4
JÂA-Net	53.8	47.8	58.2	78.5	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4
SRERL	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
UGN-B	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
HMP-PS	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
FAT	51.7	49.3	61.0	77.8	79.5	82.9	86.3	67.6	51.9	63.0	43.7	56.3	64.2
PIAP	54.2	47.1	54.0	79.0	78.2	86.3	89.5	66.1	49.7	63.2	49.9	52.0	64.1
Ours	53.3	47.4	56.2	79.4	80.7	85.1	89.0	67.4	55.9	61.9	48.5	49.0	64.5

Table 6. F1 score of supervised AU recognition on the DISFA database.

Methods	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	**Avg.**
DRML	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
ROI	41.5	26.4	66.4	50.7	8.5	89.3	88.9	15.6	48.5
SRERL	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
JÂA-Net	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
UGN-B	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS	38.0	45.9	65.2	50.9	50.8	76.0	93.3	67.6	61.0
FAT	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	61.5
PIAP	50.2	51.8	71.9	50.6	54.5	79.7	94.1	57.2	63.8
Ours	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5

work that can learn AU-related local representations, significantly enhancing performance. The local features learned by our method are more powerful than those learned by previous self-supervised methods.

5.4. Comparison with semi-supervised methods

This section evaluates whether the learned representation enables data-efficient AU recognition. On the two databases, a certain percentage of samples (i.e.., 10%, 20%, 30%, 40%, or 50%) is missed while training AU classifiers (fixed the parameters of $f(\cdot)$). Our method is compared to DAU-R [27] and BGCS [25]. BGCS is retrained because it does not provide experimental results on two databases, while the results of DAU-R on the BP4D database are directly cited. Figure 4 shows the results on two databases.

First, as the missing rate increases, the performance of the methods shows a downward trend. More supervisory information improves AU classifier training. Secondly, our results achieve considerable improvement over DAU-R and BGCS under different missing rates. Both DAU-R and BGCS summarize label distribution from limited ground-truth AU labels to constrain unlabeled data. The better performance of our method demonstrates that it can learn powerful patterns from many unlabeled images, enabling more data-efficient AU recognition than previous semi-supervised methods.

5.5. Comparison with supervised methods

We also compare our method to state-of-the-art supervised methods, including DRML [30], ROI [15], SRERL [14], JÂA-Net [22], UGN-B [23], HMP-PS [24], PIAP [26], and FAT [11]. Tables 5 and 6 show the results. In this section, we further finetune the parameters of $f(\cdot)$.

Our results outperform recent supervised methods on both databases. Our F1 score is 16.2%, 8.1%, 2.1%, 1.6%, 1.2%, 1.1%, 0.3%, and 0.4% better than DRML, ROI, JÂA-Net, SRERL, UGN-B, HMP-PS, FAT, and PIAP on the BP4D database. On the DISFA database, the result of our method is also higher than the supervised methods. Though the works try to divide face to extract AU-related features, they need fully AU-labeled images for training. Our knowledge-driven method is able to learn local patterns from large amounts of unlabeled facial images. These results demonstrate that the learned unsupervised representations are powerful and well generalized.

6. Conclusion

In this paper, we propose a novel knowledge-driven self-supervised representation learning framework for AU recognition. AU labeling rules are summarized and leveraged to guide the design of the framework. Specifically, facial areas are divided into eight separate parts according to AU labeling rules. A contrastive learning component based on the differences between facial areas is introduced to train the feature encoder. The correspondence between facial areas is also explored by a predicting learning component to enhance the representation learning. The framework is trained on a large unlabeled database. Evaluation on two benchmark databases demonstrates that the learned features outperform other self-supervised methods and have better generalization and data efficiency for AU recognition.

Acknowledgments

This work was supported by the National Key R & D program of China (2018YFB1307102) and the National Natural Science Foundation of China (92048203).

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 3, 4, 7

[2] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017. 1, 2

[3] Paul Ekman, Joseph C Hager, and Wallace V Friesen. The symmetry of emotional and deliberate facial actions. *Psychophysiology*, 18(2):101–106, 1981. 4

[4] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3, 1978. 1

[5] Shizhong Han, Zibo Meng, Zhiyuan Li, James O'Reilly, Jie Cai, Xiaofeng Wang, and Yan Tong. Optimizing filter size in convolutional neural networks for facial action unit recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5070–5078, 2018. 1, 2

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3, 7

[7] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[9] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. 3

[10] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017. 6

[11] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2, 8

[12] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 3, 4

[13] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 6

[14] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. *arXiv preprint arXiv:1904.09939*, 2019. 1, 2, 8

[15] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2017. 1, 2, 8

[16] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019. 1, 2, 3, 7

[17] Chen Ma, Li Chen, and Junhai Yong. Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. *Neurocomputing*, 355:35–47, 2019. 1

[18] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 2017. 2

[19] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 6

[20] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. *NIPS*, 2019. 2

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[22] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: Joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129(2):321–340, 2021. 2, 8

[23] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, 2021. 2, 8

[24] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6267–6276, 2021. 2, 8

[25] Yale Song, Daniel McDuff, Deepak Vasisht, and Ashish Kapoor. Exploiting sparsity and co-occurrence structure for action unit recognition. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015. 1, 2, 8

[26] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12899–12908, 2021. 2, 8

[27] Shan Wu, Shangfei Wang, Bowen Pan, and Qiang Ji. Deep facial action unit recognition from partially labeled data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3951–3959, 2017. 1, 2, 8

[28] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A

high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013. 6

[29] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. 6

[30] Kaili Zhao, Wen Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Computer Vision & Pattern Recognition*, 2016. 1, 2, 8