Supervised Contrastive Learning with Identity-Label Embeddings for Facial Action Unit Recognition

Tangzheng Lian
lian.tangzheng@kcl.ac.uk
David Adama
David.adama@ntu.ac.uk
Pedro Machado
pedro.machado@ntu.ac.uk
Doratha Vinkemeier
doratha.vinkemeier@ntu.ac.uk

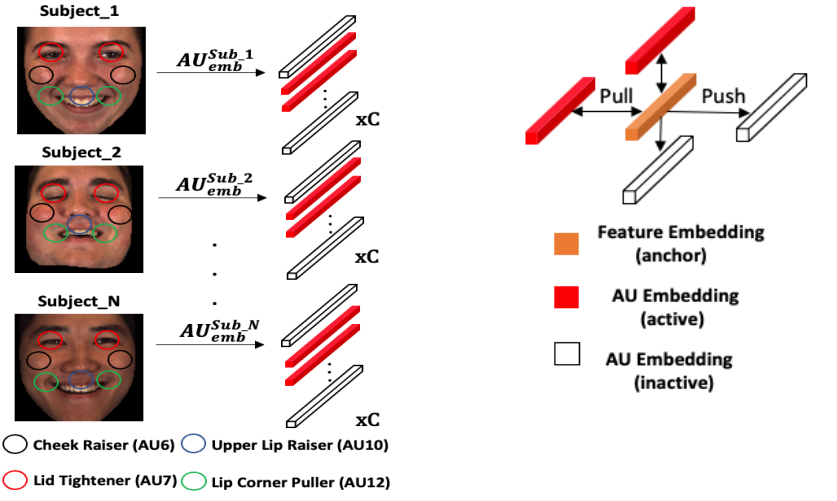
Department of Computer Science
Computing and Informatics Research Centre
Nottingham Trent University
Nottingham, UK

Abstract

Facial expression analysis is a crucial area of research for understanding human emotions. One important approach to this is the automatic detection of facial action units (AUs), which are small, visible changes in facial appearance. Despite extensive research, automatic AU detection remains a challenging computer vision problem. This paper addresses two central difficulties: the first is the inherent differences in facial behaviour and appearance across individuals, which leads current AU recognition models to overfit subjects in the training set and generalize poorly to unseen subjects; the second is representing the complex interactions among different AUs. In this paper, we propose a novel two-stage training framework, called CL-ILE, to address these long-standing challenges. In the first stage of CL-ILE, we introduce identity-label embeddings (ILEs) to train an ID feature encoder capable of generating person-specific feature embeddings for input face images. In the second stage, we present a data-driven method that implicitly models the relationships among AUs using contrastive loss in a supervised setting while eliminating the person-specific features generated by the first stage to enhance generalizability. By removing the ID feature encoder and ILEs from the first stage after training, CL-ILE becomes more lightweight and readily applicable to real-world applications than models using graph-based structures. We evaluate our approach on two widely-used AU recognition datasets, BP4D and DISFA, demonstrating that CL-ILE can achieve state-of-the-art performance on the F1 score.

1 Introduction

The automatic detection of human facial expressions is a significant area in affective computing with diverse applications including Human-Computer Interaction [4, 33], psychology [1, 19], and medicine [9, 17]. One widely studied facial expression descriptor is the Facial Action Coding System (FACS) [13]. FACS comprises a small set of visible Action



(a) (b)

Figure 1: a) Illustration of the idea of ILEs. Examples of face images are sampled from the BP4D dataset with the same AU annotations but from different subjects; b) Illustration of the implementation of contrastive loss in the second training stage of CL-ILE.

Units (AUs) that can be combined to characterize a broad range of facial expressions rendering it a long-established research topic in both affective computing and computer vision.

Gender, age and individual variations in facial shape significantly influence the detected intensity and occurrence of AUs [7], thereby explaining why existing AU classifiers frequently exhibit suboptimal performance when tested on subjects not included in the training set. Expanding the training dataset may seem like an intuitive solution to this challenge, but the labour-intensive nature of expert AU annotation makes it unfeasible. Although previous methods have endeavoured to mitigate individual bias during training, they typically necessitate a time-consuming multi-task learning framework [38, 42] and supplementary training data [37, 38].

Besides gaining insights into AU discriminative features, comprehending the relationships among AUs can offer valuable information for AU recognition [34]. For instance, a smile often involves the co-occurrence of AU6 (Cheek Raise) and AU12 (Lip Corner Puller), whereas AU1 (Inner Brow Raiser) and AU4 (Brow Lowerer) are mutually exclusive based on their definitions and visual characteristics. Therefore, detecting AUs is often viewed as a multi-label classification task. Although some studies have explored utilizing AU co-occurrence by designing network structures such as Graph Convolutional Networks (GCN) [20, 27, 28] or general Graph Neural Networks (GNN) [24], these approaches often entail high computational costs. Other works concentrate on learning AU relationships based on facial regions [14, 31], but they do not explore the correlations in AU annotations.

In this paper, we introduce CL-ILE, a two-stage training framework for AU recognition in a supervised setting that incorporates identity-label embeddings (ILEs) and contrastive loss. In the first training stage, we augment AU label embeddings by incorporating an additional dimension representing the subject index, resulting in the creation of ILEs. As shown in Fig. 1(a), each subject is associated with a unique set of AU embeddings. Thus, even when the input face images have identical AU annotations, the AU embeddings will vary based on the

respective identities. In the second stage, as illustrated in Fig. 1(b), we employ contrastive loss, where the feature embedding serves as the anchor, pushing away from inactive AU embeddings while pulling towards active AU embeddings during training to implicitly learn the correlations among AUs. During the process of learning AU relationships, we mitigate the influence of the ID feature embedding generated by the ID feature encoder in the first stage by minimizing its similarity to the feature embedding produced by the feature encoder in the second stage. The contributions of our work can be summarized as follows: (i) We present a novel and effective approach for addressing individual bias in AU detection that exhibits improved generalizability across subjects (ii) We utilize contrastive loss to learn AU label correlations in a data-driven manner during training, providing a simpler yet more powerful alternative to certain graph-based methods. (iii) We demonstrate the effectiveness of our approach compared to state-of-the-art methods on the BP4D and DISFA datasets.

2 Related work

In this section, we will discuss the current methods in identity-aware AU recognition, AU correlation modelling and the development of contrastive learning.

Identity-aware AU recognition. Research in identity-aware AU recognition can be divided into two main branches. The first branch involves learning person-specific models for test set subjects by transferring knowledge from the training set. Wang *et al*. [40] proposed a personalized generative adversarial recognition network for recognizing multiple facial action units by transferring facial images from a source domain to a target domain. Likewise, Chu *et al*. [7] proposed the Selective Transfer Machine, which trains personalized Support Vector Machine (SVM) classifiers for individual subjects. However, those methods need to be fine-tuned every time a new subject is added. The second branch aims to develop AU detection systems capable of learning generalized AU features by mitigating individual bias, thus avoiding overfitting the training set. Zhang *et al*. [42] proposed an adversarial training framework (ATF) that minimizes the AU loss while maximizing the identity (ID) loss to reduce the influence of personal identity. Tu *et al*. [38] proposed multi-task network cascades with the specialized face clustering task to decrease individual bias using translation transformation in the feature space. However, these methods can be computationally expensive due to their reliance on multi-task learning. Another approach is LP-Net [31], which uses normalized facial landmarks as person-specific features and attempts to eliminate their impact by making global AU features orthogonal to them. This method assumes that facial landmarks can accurately represent person-specific features, which may not always hold true. Our method was inspired by the anti-person-specific module proposed in PIAP-DF [37]. They addressed this issue by introducing a feature encoder that generates person-specific features and makes the general feature encoder orthogonal to it. In contrast, our method introduces identity-label embeddings that explicitly ensure the model learns person-specific features, even when subjects have the same AU annotation, without using additional training data.

AU correlation modelling. AU recognition is often treated as a multi-label classification task due to the co-occurrence of AUs. Various methods have been proposed to leverage the relationships among AUs to enhance performance. Graph-based methods such as GCN [20, 27, 28] and general GNN [24] have been proposed to explicitly model these relationships. However, these methods can be computationally demanding as they require an additional network for AU relationship modelling. Alternatively, some researchers have used

AU correlation statistics calculated from the AU recognition dataset as prior knowledge for training [29, 43], but they only considered pairwise AU correlations. Chen *et al*. [6] proposed a supervised hierarchical contrastive learning framework to model three kinds of correlations among AUs (i.e., unary, binary, and multivariate) and they treat each AU with the same label from different subjects as positive samples in a minibatch and pull their representations together. However, this approach may not be intuitive as it forces features from different subjects to be consistent. Instead, we introduce an additional training stage in advance to train an encoder that generates person-specific features with ILEs, which are then eliminated. Furthermore, our proposed AU correlation learning module does not include additional network structures and can implicitly learn correlations across multiple AUs during training.

Contrastive learning is a cutting-edge technique for representation learning, initially developed in the context of unsupervised or self-supervised learning [15, 18]. The fundamental idea of contrastive loss is to define an anchor sample, positive samples, and negative samples, and then optimize the embedding space by minimizing the distance between the anchor sample and positive samples while maximizing the distance between the anchor sample and negative samples. Recently, SupCon [22] was proposed to apply contrastive loss in a supervised setting in image classification by choosing one image-level representation of an image as the anchor and pushing the representation of the anchor close to its positive samples (e.g., augmentations of the anchor or images with the same label as the anchor in the minibatch). However, SupCon is specifically designed for single-label image classification and is not directly applicable to multi-label classification tasks, such as AU recognition. The challenge of adapting SupCon to multi-label classification lies in defining positive and negative samples, as images in this task are annotated with multiple labels, which are often highly correlated. Our approach is inspired by [2], which successfully leverages contrastive loss in multi-label classification and captures the label correlations to enhance the predictive performance. However, our model differs from theirs in terms of encoder structures and includes an identity learning module, which is tailored for AU recognition.

3 Methodology

3.1 Overview

Given a facial image $I \in \mathbb{R}^{3\times H\times W}$, the goal of AU recognition is to predict the probabilities of each AU occurring in the image. This task is a multi-label classification problem since multiple AUs can coexist in the same image. Therefore, the output of an AU recognition model can be represented as a vector of probabilities $P = \{\hat{p}^c \in \mathbb{R}\}_{c=1}^{C}$, where C is the number of AU classes and $\hat{p}^c$ represents the probability of the c-th AU occurring in the image.

Our proposed method, CL-ILE, aims to mitigate inherent differences among individuals while capturing intricate correlations among AUs. CL-ILE comprises two training stages, as illustrated in Fig. 2. The first stage primarily focuses on training an ID feature encoder, F_{enc}^{ID}, which generates person-specific features, F_{emb}^{ID}, from the input facial image. In contrast, the second stage aims to train a general feature encoder, F_{enc}^{G}, by minimizing the similarity between F_{emb}^{G} and F_{emb}^{ID} during training.

For each facial image with a subject ID in the batch, CL-ILE initially processes it through the backbone network for feature extraction. In the first training stage, CL-ILE utilizes the feature maps produced by the backbone to train F_{enc}^{ID} and ILEs. In the second training stage, both subject ID and ILEs are discarded since we only require F_{enc}^{ID} for generating

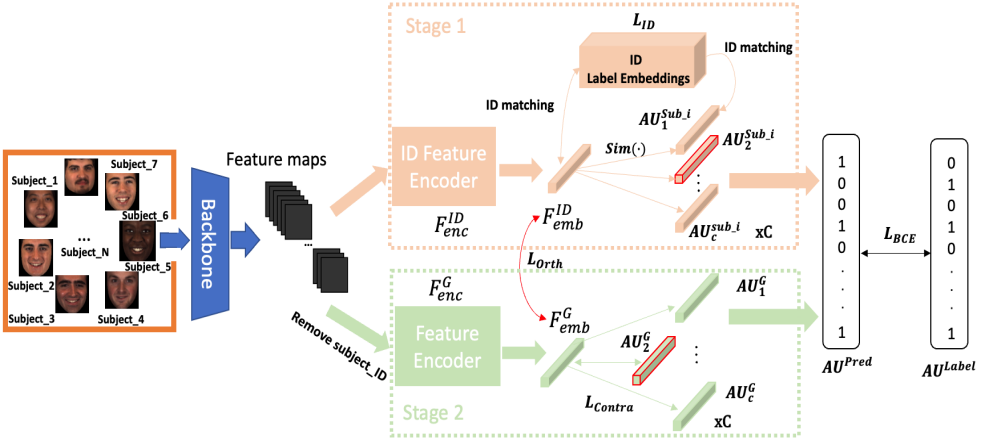


Figure 2: An overview of our proposed framework.

F^{ID}_{emb}. Following the completion of the second training stage, F^{ID}_{enc} is also discarded, and our approach produces F^{G}_{enc} and a set of AU embeddings that demonstrate enhanced generalizability and capture AU correlations. These are subsequently employed for making predictions during validation. For the underlying network architecture in CL-CLE, we have chosen ResNet18 [17] with the last pooling and fully connected layers removed. The resulting feature maps are utilized as feature representations for subsequent encoding. The output of the backbone network consists of 512 feature maps, each measuring 7x7 in size. These are then flattened into 49 representations, each of dimension 512. Subsequently, we input these 49 representations of the input image into a feature encoder to generate the feature-level embedding. We have found that Transformer Encoder layers [39] are both simple and effective as feature encoders. Therefore, in the second training stage, we employ Transformer Encoder layers as the network architecture for both F^{ID}_{enc} and F^{G}_{enc}, sharing the same network parameters, to generate F^{ID}_{emb} and F^{G}_{emb} in the latent space.

In contrast to traditional AU recognition models, our approach emphasizes representation learning. Consequently, the probability of the c-th AU, $\hat{p}^c$, is not generated by the final linear layer of a classification network. Instead, we compute $\hat{p}^c$ by taking the inner product of F^{G}_{emb} and AU^{c}_{emb}.

3.2 Training Stage I

In the first training stage, the objective is to obtain a feature encoder capable of generating person-specific feature embeddings. To achieve this, we introduce ID label embeddings (ILEs) as trainable parameters initialized with Gaussian distributions and normalized during training. The shape of the ILEs is $N \times C \times D$, where N represents the number of subjects in the training set, C represents the number of AU classes, and D represents the dimension of the embeddings. During this stage, we retain the subject ID of the input image to ensure the network's awareness of the identities. Once we obtain F^{ID}_{emb} from F^{ID}_{enc}, we search the ILEs to retrieve the set of AU embeddings corresponding to the identity of the input image. The goal in this stage is to learn distinct person-specific features. Thus, we propose an ID loss that encourages the aggregation of different AU embeddings belonging to the same subject

while pushing apart AU embeddings of the same kind associated with different subjects in ILEs during training. The ID loss can be formulated as:

$$\mathcal{L}_{ID} = \log \frac{\sum_{i=1}^{N}\sum_{j=1}^{C} |sim(ILE_{i,j}, \bar{ILE}_i)|}{\sum_{j=1}^{C}\sum_{i=1}^{N} |sim(ILE_{i,j}, \bar{ILE}_j)|} \tag{1}$$

where the $sim(\cdot)$ function represents cosine similarity, $\bar{ILE}_i$ represents the mean across all AUs of subject i, $\bar{ILE}_j$ represents the mean of AU j across all identities, and $ILE_{i,j}$ represents the feature vector of a specific AU j for the subject i . Additionally, we apply the widely used Binary Cross Entropy (BCE) loss function for AU classification. The BCE loss can be expressed as:

$$\mathcal{L}_{BCE} = -\frac{1}{C}\sum_{c=1}^{C} w_c \left[y_c \log s\left(F_{emb}^{ID} \cdot AU_c^{Sub_i}\right) + (1-y_c)\log\left(1 - s\left(F_{emb}^{ID} \cdot AU_c^{Sub_i}\right)\right)\right] \tag{2}$$

Here, y_c denotes the ground truth label for the c-th AU, w_c is the weight calculated by Selective Learning [16] using the uniform distribution and $s(\cdot)$ represents the sigmoid function. Consequently, the final loss for the first training stage can be expressed as:

$$\mathcal{L}_1 = \lambda \mathcal{L}_{BCE} + (1-\lambda)\mathcal{L}_{ID} \tag{3}$$

Here, λ is a hyperparameter that controls the weight between $\mathcal{L}_{ID}$ and $\mathcal{L}_{BCE}$.

3.3 Training Stage II

In the second training stage, the backbone and feature encoder networks retain the same structure and parameters as in the first stage. However, during training at this stage, we do not retain the subject ID. Instead, we freeze and save the network parameters from stage one, utilizing only the output F_{emb}^{ID} during stage two training. The objective in this stage is to eliminate person-specific features and acquire a generalized model. To achieve this, we apply $\mathcal{L}_{Orth}$ from Eq. (4) during training to reduce the cosine similarity between F_{emb}^{G} and F_{emb}^{ID}, where B is the batch of samples.

$$\mathcal{L}_{Orth} = \frac{1}{|B|}\sum_{(F_{emb}^{ID}, F_{emb}^{G})\in B} \log|sim(F_{emb}^{ID}, F_{emb}^{G})| \tag{4}$$

Moreover, we have introduced contrastive loss [2, 5] in a supervised setting to model AU correlations. The central idea is that AU embeddings should exhibit proximity when their AU labels frequently co-occur and distance when their AU labels seldom co-occur. The contrastive loss encourages the clustering of correlated AU embeddings and the separation of unrelated ones. In the multi-label AU recognition scenario, we designate the feature embedding as the anchor sample, positive AU embeddings as the positive samples, and negative AU embeddings as the negative samples. The $\mathcal{L}_{Contra}$ can be formulated as follows:

$$\mathcal{L}_{Contra} = \frac{1}{|B|}\sum_{(F_{emb}^{G}, y)\in B} \frac{1}{|P(y)|}\sum_{p\in P(y)} -\log\frac{exp(sim\left(F_{emb}^{G}, AU_p^{G}\right)/\tau)}{\sum_{c=1}^{C} exp(sim\left(F_{emb}^{G}, AU_c^{G}\right)/\tau)} \tag{5}$$

where $P(y)$ represents the subset of positive AU labels, AU_p^G denotes the p-th AU embedding and τ is the scaling parameter . The loss function for the second training stage combines the $\mathcal{L}_{BCE}$ from Eq. (2), $\mathcal{L}_{Orth}$, and $\mathcal{L}_{Contra}$, as follows:

$$\mathcal{L}_2 = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \mathcal{L}_{Orth} + \lambda_3 \mathcal{L}_{Contra} \tag{6}$$

where λ_1, λ_2 and λ_3 are weights to balance different losses.

4 Experiments

4.1 Implementation Details

Our framework is trained in two stages. Before the two stages, we pre-trained our backbone ResNet18 with weights from ImageNet [23] and saved it as the feature extractor. The weights of the backbone remain frozen during both training stages. Upon completing stage one, we save and freeze F_{enc}^{ID} and utilize it to obtain F_{emb}^{ID}. During stage two training, we discard F_{enc}^{ID}, preserve F_{emb}^{G} and AU embeddings, and utilize them for predictions on unseen subjects during validation. We employed SGD as the optimizer for all networks, using an initial learning rate of 1e-4 for the backbone, 1e-5 for both training stages on BP4D and 5e-5 on DISFA. The learning rate decreased when the F1 score reached a plateau. A batch size of 64 was used for both BP4D and DISFA. λ was set to 0.4 empirically to balance losses in stage one, while in stage two, λ_1 was set to 1, λ_2 to 0.5, and λ_3 to 0.4. All embeddings in our method have a consistent dimension of 512. The scaling parameter τ in the contrastive loss is set to 2. The threshold in the sigmoid function for binary classification was set to 0.5. The backbone was trained for 300 epochs, while CL-ILE was trained for 20 epochs on BP4D and 150 epochs on DISFA. All implementations were conducted using PyTorch [32], and the models were trained and evaluated on an NVIDIA 3090Ti GPU.

4.2 Datasets&Metrics:

We assess CL-ILE using two prevalent AU datasets, BP4D [41] and DISFA [30]. BP4D is a spontaneous AU dataset comprising 23 female and 18 male adult participants. Each participant completed eight sessions with distinct target emotions, and videos from each session were recorded, accompanied by frame-level binary AU occurrence labels. Altogether, around 140,000 frames are annotated with AUs. DISFA contains videos of 27 subjects with diverse genders and ethnicities, all of whom viewed a 4-minute video while their facial reactions were recorded by two cameras on the right and left. In total, approximately 130,000 frames are annotated with discrete AU intensities ranging from 0 to 5. Intensities equal to or greater than 2 are considered positive. We adhere to the subject-exclusive 3-fold cross-validation protocol from previous studies [6, 21, 24, 25, 27, 28, 31, 35, 37, 42] and select 8 and 12 AUs that frequently appear in DISFA and BP4D, respectively, for evaluation.

We employ CE-CLM [3] to detect facial landmarks and conduct facial alignment as the image pre-processing step for both datasets. Subsequently, we normalize and resize all images to 240 x 240 pixels. During training, we apply random cropping (224x224) and random horizontal flipping for data augmentation. In our study, we use the prevalent frame-level F1 score in AU recognition, defined as $F1 = 2\frac{P \cdot R}{P+R}$, to ensure a standardized comparison with previous works. Additionally, **Avg.** signifies the unweighted mean over all AUs: 12 in BP4D and 8 in DISFA.

Method/AUs	1	2	4	6	7	10	12	14	15	17	23	24	Avg.
DRML† [40]	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
ATF* [18]	39.2	35.2	45.9	71.6	71.9	79.0	83.7	65.5	33.8	60.0	37.3	41.8	55.4
LP-Net*† [31]	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
AU-GCN† [27]	46.8	38.5	60.1	**[80.1]**	[79.5]	84.8	88.0	67.3	52.0	63.2	40.9	52.8	62.8
SRERL† [24]	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
PIAP† [32]	[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
SupHCL*† [1]	52.8	45.7	**[61.6]**	79.5	79.3	84.7	86.9	[67.6]	51.4	62.5	48.6	52.3	64.4
ME-Graph† [33]	52.7	44.3	[60.9]	[79.9]	**[80.1]**	[85.3]	**[89.2]**	**[69.4]**	**[55.4]**	**[64.4]**	[49.8]	[55.1]	**[65.5]**
CL-ILE (Ours)	**[55.1]**	**[52.1]**	55.0	78.2	75.5	83.4	[88.1]	67.4	51.9	59.5	46.9	**[62.2]**	[64.6]

Table 1: Comparisons of F1 scores (in %) achieved by state-of-the-art methods and ours for 12 AUs on the BP4D dataset, where the methods that are identity-aware are marked by * while the methods that dive into AU relationship modelling are marked by †. The best and second-best results of each column are indicated in bold font with brackets and brackets only, respectively.

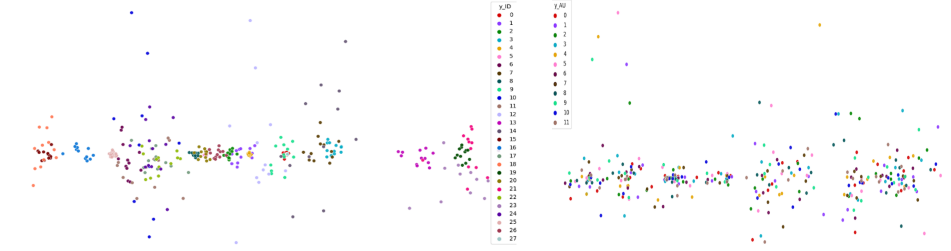
Method/AUs	1	2	4	6	9	12	25	26	Avg.
DRML† [40]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
ATF* [18]	45.2	39.7	47.1	48.6	32.0	55.0	86.4	39.2	49.2
AU-GCN† [27]	32.3	19.5	55.7	[57.9]	**[61.4]**	62.7	90.9	[60.0]	55.0
SRERL† [24]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
LP-Net*† [31]	29.9	24.7	[72.7]	46.8	49.6	72.9	[93.8]	**[65.0]**	56.9
ME-Graph† [33]	[54.6]	47.1	**[72.9]**	54.0	[55.7]	[76.7]	91.1	53.0	[63.1]
PIAP† [32]	50.2	[51.8]	71.9	50.6	54.5	**[79.7]**	**[94.1]**	57.2	**[63.8]**
CL-ILE (Ours)	**[58.9]**	**[56.4]**	69.1	**[58.5]**	54.4	72.2	85.9	47.3	62.8

Table 2: Comparisons of F1 scores (in %) achieved by state-of-the-art methods and ours for 8 AUs on the DISFA dataset, where the methods that are identity-aware are marked by * while the methods that dive into AU relationship modelling are marked by †. The best and second-best results of each column are indicated in bold font with brackets and brackets only, respectively.

4.3 Backbone Selection

Feature extraction serves as the fundamental component of AU recognition, typically accomplished through a backbone network. In our endeavour to identify the most efficient backbone for AU recognition, we subjected a series of widely recognised architectures to rigorous testing using the BP4D dataset. We strictly adhered to a 3-fold cross-validation protocol to ensure the reliability of our results. From our evaluations, as summarized in Table 4, the top five performing backbone architectures were identified. Among these, ResNet18 distinguished itself by consistently demonstrating superior performance metrics, and as a result, was chosen as the backbone for CL-ILE.

Moving beyond conventional CNN architectures, we also explored the potential of several state-of-the-art transformer-based models such as CoAtNet [10], ViT [11], and Twins-SVT [8]. However, these models, despite their growing popularity in other domains, produced an F1 score of less than 60.0% in our tests. One possible reason for this underperformance might be the inherent design of transformer-based methods. These models are often tailored for handling large-scale datasets and extracting global features. However, they might not be adept at finely capturing the localized features in images, where the AUs predominantly manifest. Consequently, CNN-based architectures, especially those like ResNets and EfficientNets, continue to dominate as the preferred choices for AU recognition tasks, with the noteworthy exception of Swin-Transformer [26].



(a) (b)

Figure 3: PCA visualization of the ILEs on BP4D (fold 1): a) Each colour represents a subject in the training set and the number of dots for each colour represents the AUs accordingly; b) Each colour represents an AU and the number of dots for each colour represents the subjects in the training set accordingly.

	Backbone	L_{ID}	L_{Orth}	L_{Contra}	**Avg.**
A_0	✓				61.8
A_1	✓			✓	63.0
A_2	✓		✓		62.7
A_3	✓		✓	✓	63.7
A_4	✓	✓	✓		64.3
CL-ILE	✓	✓	✓	✓	64.6

Table 3: Ablation experiments on BP4D.

Arch.	**Avg.**
ResNet18	61.8
Swin-Transformer [26]	61.7
ResNet152	61.3
ResNet50	61.1
EfficientNet-B2 [36]	61.0

Table 4: Backbone selection on BP4D.

4.4 Ablation Study

The effectiveness of each component of our framework is assessed in Table 3. To decipher the contributions of each module, we defined specific stages within our ablation studies, represented as A_0 through A_4, which correspond to configurations with or without certain integral modules. The model denoted as A_1 in the table refers to the absence of L_{Orth}, indicating that only the second training stage and contrastive loss are employed. In this scenario, A_1 attains an average F1 score of 63.0%, representing a 1.2% improvement over the baseline A_0 solely by considering the AU correlations. By introducing the first training stage and L_{ID}, we observe enhancements of 0.9% and 1.6% when comparing A_2 to A_0 and A_4 to A_2, respectively.

Additionally, as illustrated in Fig. 3(a) and Fig. 3(b), there is a clear distinction in data distribution patterns with ILEs. In Fig. 3(a), dots of the same colour predominantly form cohesive clusters, suggesting strong intra-subject similarities. Conversely, in Fig. 3(b) same colored dots display a more dispersed pattern, indicating increased inter-subject variance. These visual observations reinforce the effectiveness of our proposed ILEs and L_{ID}. Likewise, in Fig. 4, the substantial resemblance between the correlation matrix of the ground truth and CL-ILE demonstrates the effectiveness of our framework in modelling the correlations among AUs. For detailed insights into the derivation of the correlation matrix and training specifics of parameters λ, λ_1, λ_2, and λ_3, readers are directed to our supplementary materials.

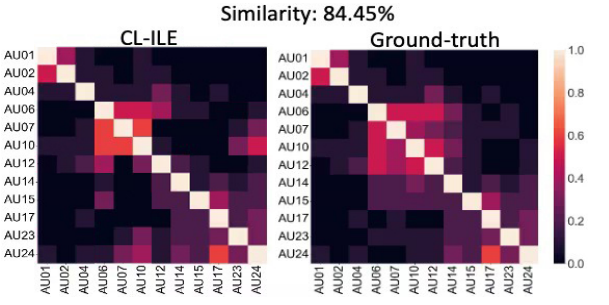


Figure 4: The correlation matrix of the CL-CLE and ground truth based on BP4D.

4.5 Comparison to the State-of-the-arts

The performance of our method in AU recognition, as illustrated in Tables 1 and 2, surpasses existing benchmarks. Notably, CL-ILE ranks second in BP4D and exhibits strong competitiveness on DISFA. Our model outperforms state-of-the-art techniques on BP4D by a margin of 0.9% in AU1, 5.0% in AU2, and 7.1% in AU24. Regarding DISFA, it also excels beyond existing methodologies by 4.3% in AU1, 5.4% in AU2, and 0.6% in AU6. In terms of AU relation modelling alone, model A_1 (in Table 3) reaches 63.0% on BP4D, surpassing DRML [21], AU-GCN [27], and SRERL [24] which delve into AU relation modelling. Furthermore, in the context of identity-aware AU recognition, model A_4 (in Table 3) scores 64.3% on BP4D, outperforming LP-Net [31] and PIAP [37], which focus on identity-aware AU recognition. Lastly, our method is not only more lightweight but also outperforms several Graph Neural Network (GNN) based approaches like AU-GCN [27] and SRERL [24].

5 Conclusion

In conclusion, we present a novel two-stage training framework, CL-ILE, which efficiently handles the complex relationships among AUs and the inherent differences in facial behaviour across individuals. By introducing identity-label embeddings (ILEs) in the first stage and employing a data-driven method with contrastive loss in the second stage, CL-ILE effectively enhances the generalizability of AU recognition models. The removal of the ID feature encoder and ILEs after training further improves the model's lightweight nature, making it more suitable for real-world applications compared to graph-based structures. Evaluations on two widely-used AU recognition datasets, BP4D and DISFA, demonstrate that our approach can achieve state-of-the-art performance on the F1 score, signifying its potential to contribute significantly to the field of facial expression analysis and understanding human emotions.

While our research presents valuable insights, it also acknowledges certain areas for improvement. The datasets we used, though respected in the field, may not capture the full spectrum of individual variability. In addition, our study did not encompass cross-dataset evaluations due to time considerations. These aspects provide exciting directions for our future endeavours. We are optimistic about delving deeper into these areas in our subsequent studies.

References

[1] Alex A Ahmed and Matthew S Goodwin. Automated detection of facial expressions during computer-assisted instruction in individuals on the autism spectrum. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6050–6055, 2017.

[2] Junwen Bai, Shufeng Kong, and Carla P Gomes. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *International Conference on Machine Learning*, pages 1383–1398. PMLR, 2022.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. doi: 10.1109/FG.2018.00019.

[4] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *2003 Conference on computer vision and pattern recognition workshop*, volume 5, pages 53–53. IEEE, 2003.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Yingjie Chen, Chong Chen, Xiao Luo, Jianqiang Huang, Xian-Sheng Hua, Tao Wang, and Yun Liang. Pursuing knowledge consistency: Supervised hierarchical contrastive learning for facial action unit recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 111–119, 2022.

[7] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3515–3522, 2013.

[8] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021.

[9] UmurAybars Ciftci, Xing Zhang, and Lijun Tin. Partially occluded facial action recognition and interaction in virtual reality applications. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 715–720. IEEE, 2017.

[10] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Joy Egede, Michel Valstar, and Brais Martinez. Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 689–696, 2017. doi: 10.1109/FG.2017.87.

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

[14] Xuri Ge, Pengcheng Wan, Hu Han, Joemon M. Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Local global relational network for facial action units recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08, 2021. doi: 10.1109/FG52635.2021.9666961.

[15] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

[16] Emily Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

[18] 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.

[19] Rens Hoegen, Jonathan Gratch, Brian Parkinson, and Danielle Shore. Signals of emotion regulation in a social dilemma: Detection from face and context. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7, 2019. doi: 10.1109/ACII.2019.8925478.

[20] Xibin Jia, Yuhan Zhou, Weiting Li, Jinghua Li, and Baocai Yin. Data-aware relation learning-based graph convolution neural network for facial action unit recognition. *Pattern Recognition Letters*, 155:100–106, 2022.

[21] Zhao Kaili, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[24] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang 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.

[25] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun 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 (FG 2017)*, pages 103–110. IEEE, 2017.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[27] Zhilei Liu, Jiahui Dong, Cuicui Zhang, Longbiao Wang, and Jianwu Dang. Relation modeling with graph convolutional networks for facial action unit detection. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 489–501. Springer, 2020.

[28] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1239–1246, 2022.

[29] 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.

[30] 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.

[31] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11917–11926, 2019.

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[33] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W. Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19):eaao6760, 2018. doi: 10.1126/scirobotics.aao6760.

[34] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 13(3):1274–1289, 2019.

[35] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaa-net: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321–340, 2021.

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[37] 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.

[38] Cheng-Hao Tu, Chih-Yuan Yang, and Jane Yung-jen Hsu. Idennet: Identity-aware facial action unit detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Can Wang and Shangfei Wang. Personalized multiple facial action unit recognition through generative adversarial recognition network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 302–310, 2018.

[41] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10): 692–706, 2014.

[42] Zheng Zhang, Shuangfei Zhai, Lijun Yin, et al. Identity-based adversarial training of deep cnns for facial action unit recognition. In *BMVC*, page 226. Newcastle, 2018.

[43] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2015.