International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02348-z

Sample-Cohesive Pose-Aware Contrastive Facial Representation Learning

Yuanyuan Liu[1] · Shaoze Feng[1] · Shuyang Liu[1] · Yibing Zhan[2] · Dapeng Tao[2,3] · Zijing Chen[4] · Zhe Chen[4]

Received: 29 March 2024 / Accepted: 6 January 2025

Abstract

Self-supervised facial representation learning (SFRL) methods, especially contrastive learning (CL) methods, have been increasingly popular due to their ability to perform face understanding without heavily relying on large-scale well-annotated datasets. However, analytically, current CL-based SFRL methods still perform unsatisfactorily in learning facial representations due to their tendency to learn pose-insensitive features, resulting in the loss of some useful pose details. This could be due to the inappropriate positive/negative pair selection within CL. To conquer this challenge, we propose a Pose-disentangled Contrastive Facial Representation Learning (PCFRL) framework to enhance pose awareness for SFRL. We achieve this by explicitly disentangling the pose-aware features from non-pose face-aware features and introducing appropriate sample calibration schemes for better CL with the disentangled features. In PCFRL, we first devise a pose-disentangled decoder with a delicately designed orthogonalizing regulation to perform the disentanglement; therefore, the learning on the pose-aware and non-pose face-aware features would not affect each other. Then, we introduce a false-negative pair calibration module to overcome the issue that the two types of disentangled features may not share the same negative pairs for CL. Our calibration employs a novel neighborhood-cohesive pair alignment method to identify pose and face false-negative pairs, respectively, and further help calibrate them to appropriate positive pairs. Lastly, we devise two calibrated CL losses, namely calibrated pose-aware and face-aware CL losses, for adaptively learning the calibrated pairs more effectively, ultimately enhancing the learning with the disentangled features and providing robust facial representations for various downstream tasks. In the experiments, we perform linear evaluations on four challenging downstream facial tasks with SFRL using our method, including facial expression recognition, face recognition, facial action unit detection, and head pose estimation. Experimental results show that PCFRL outperforms existing state-of-the-art methods by a substantial margin, demonstrating the importance of improving pose awareness for SFRL. Our evaluation code and model will be available at https://github.com/fulaoze/CV/tree/main.

Keywords Self-supervised facial representation learning · False-negative pair calibration · Facial expression analysis

1 Introduction

Learning facial representations is an important task in computer vision. With the ability to analyze faces, we can obtain

Communicated by Boxin Shi.

✉ Zhe Chen
zhe.chen@latrobe.edu.au

Yuanyuan Liu
liuyy@cug.edu.cn

Shaoze Feng
fengshaoze@cug.edu.cn

Shuyang Liu
shuyangliu@cug.edu.cn

Yibing Zhan
zhanybjy@gmail.com

Dapeng Tao
dapeng.tao@gmail.com

Zijing Chen
zijing.chen@latrobe.edu.au

[1] School of Computer Science, China University of Geosciences, Wuhan, China

[2] Yunnan United Vision Technology Co.Ltd, Kun Ming Shi, China

[3] Yunnan University, Kun Ming Shi, China

[4] Cisco-La Trobe Centre for AI and IoT, School of Computing, Engineering and Mathematical Sciences, La Trobe University, Bundoora, Australia

Published online: 28 January 2025

various information like identities, emotions, and gestures, which lead to rich applications in various domains like facial expression recognition, face recognition, human-computer interaction, head pose estimation, and emotion analysis. Recently, deep convolutional neural networks (DCNNs) (Gamble and Huang, 2020; Zhao et al., 2016) have achieved promising facial understanding results, but they heavily rely on large, well-labeled data for supervised learning, which requires substantial manual annotation efforts and may not generalize well on other datasets. Instead of using well-labelled datasets, in recent years, self-supervised learning (SSL) has emerged as a promising alternative to train visual representation models without explicit annotations.

Current self-supervised facial representation learning (SFRL) approaches (Chen et al., 2020; Li et al., 2019; He et al., 2020) widely apply contrastive learning (CL) strategy. In a typical CL-based SFRL method, researchers first leverage pre-defined data transformations to create positive and negative samples, i.e., augmentations of the same image generate positive samples, and different other images are represented as negative samples. Then, the CL-based SFRL method will pull two features representing the same type of samples closer to each other and push those of different types far away from each other (Li and Shan, 2023; Madhusudana et al., 2022), contrasting the learning on positive and negative samples. This allows models to learn meaningful visual representations from unlabeled data. Using this methodology, existing CL-based SFRL methods (Roy and Etemad, 2021; Shu et al., 2022) have achieved promising performance on learning from unlabelled face images. This method also illustrates descent generalization abilities to downstream face-related tasks (Shu et al., 2022).

Despite progress, we found that utilizing the vanilla CL-based SFRL methods could be still sub-optimal due to the variances in facial poses. Specifically, existing approaches use an instance-level positive/negative pair selection strategy, i.e., augmentations like image flipping to help generate the collection of positive sample pairs and other different face images as negative pairs. In such a manner, a model would learn pose-insensitive representations from the positive pairs, making the recognition and appropriate handling of the facial pose variances very challenging. Nevertheless, we argue that pose information is of great importance for robust facial understanding (Samanta and Guha, 2017); for example, when a person would likely lower their head when they express sadness on the face. Therefore, we propose to enhance the pose awareness for SFRL, so that the learning on both pose-aware features and non-pose face-aware features would not affect each other and facial understanding can be improved promisingly. Figure 1 shows an overview of our motivation.

By addressing the above problem for enhancing pose awareness, we propose a novel Pose-disentangled Contrastive Facial Representation Learning (PCFRL) framework in this study. In general, the PCFRL first disentangles pose-aware features from non-pose face-aware features and then introduces a more appropriate pair calibration scheme to augment the CL on both types of features. To achieve this, we devise a three-component framework. The first component is to disentangle pose-aware and non-pose face-aware features so that the learning of these features does not affect each other. With the disentangled features, we observe that the same pose or person might generate false negative sample pairs, which may introduce significant confusion for CL. For example, two images from different persons with a shared pose would be initially selected as a negative pair, which is inappropriate for CL on the pose-aware features; meanwhile, choosing two different images with the same person as a negative pair is also inappropriate for CL on the face-aware features. Therefore, to overcome the issue, the second component, namely false-negative pair calibration module, employs a novel neighborhood-cohesive pair alignment (NPA) method to further refine the pose and face positive/negative sample pair selection for contrastive learning with the disentangled features more appropriately. Lastly, we devise two calibrated CL losses for adaptive learning on the calibrated sample pairs, obtaining the third component in this framework which ultimately enhancing the pose awarenesss SFRL for downstream tasks.

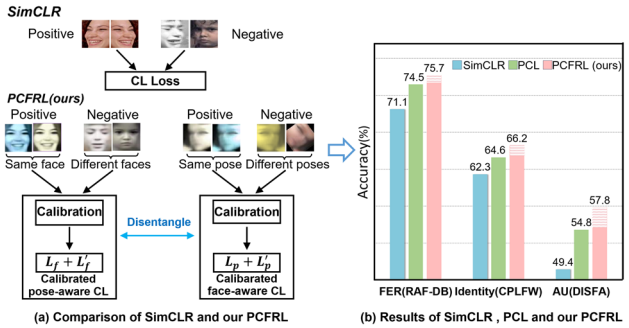


Fig. 1 As shown in (**a**), affected by different poses, the widely used contrastive learning (CL) methods, such as SimCLR, treat pose and other facial information uniformly, which may lead to suboptimal results. To alleviate this problem, our PCFRL—an extended version of our conference method (PCL)—first disentangles pose-aware and non-pose face-aware features and then calibrates face and pose false-negative pairs for more efficient calibrated pose-aware CL and calibrated face-aware CL, respectively. As shown in (**b**), our PCFRL enhances pose awareness for SFRL and improves face understanding performance promisingly

It is worth mentioning that this study is an extended version of our conference paper, PCL (Pose-disentangled Contrastive Learning) (Liu et al., 2023) published in CVPR 2023. In our original PCL paper, we primarily introduced an effective pose disentanglement algorithm for contrastive learning. In this extended study, we identified that the original PCL method suffers from inaccurate positive/negative

pait selection after disentanglement. Accordingly, we propose the NPA and calibrated CL losses to tackle this issue, which further promisingly improve our original method.

Overall, the contributions of this paper are listed as follows:

1. We propose a novel framework, PCFRL, to effectively enhance the pose awareness of SFRL. We demonstrated that enhancing pose awareness is important for robust SFRL performance.
2. Compared to our conference version (Liu et al., 2023), we introduce a false-negative pair calibration module for the SFRL with disentangled features. Specifically, we introduce an effective neighborhood-cohesive pair alignment method to help identify false negative pairs for more effective contrastive learning on both pose-aware and non-pose face-aware features.
3. Furthermore, we devise calibrated CL losses, incorporating the calibrated pairs (i.e., false-negative pairs) for calculating the contrastive loss, resulting in two new specifically calibrated pose-aware and face-aware loss functions for SFRL. These losses dynamically optimize the calibrated pairs through an adaptive weighting scheme, ultimately enhancing the learning of robust, pose-aware self-supervised facial representations.
4. We performed extensive experiments to demonstrate the significant advantages of PCFRL over our conference version method. Moreover, we also illustrate the superiority of PCFRL over existing compelling SFRL methods on several downstream tasks, including facial expression recognition (FER), facial action unit (AU) detection, facial recognition (FR), and head pose estimation (HPE), accessing state-of-the-art performance.

2 Related Work

2.1 Self-supervised Facial Representation Learning

Self-supervised facial representation learning is an important task in computer vision. It aims to automatically learn valuable features from unlabeled facial image data without relying on manually added labels. These learned features can be applied to a variety of facial-related tasks such as facial expression recognition (FER), face recognition (FR), facial action unit (AU) detection, and more (Chang et al., 2021; Jakab et al., 2018; Li et al., 2019). Zhao et al. (2015) proposes a new facial expression recognition method based on deep learning by combining deep belief networks (DBN) with multi-layer perceptrons (MLP). FAb-Net (Jakab et al., 2018) has been particularly successful in the FER task by utilizing motion variations across frames in a video to obtain facial motion features. The TwinCycle Autoencoder, proposed by Li et al. (2022, 2019), is designed to separate facial action-related movements from head movement-related movements. This separation results in robust facial emotion representations that have been demonstrated to be effective in self-supervised AU detection. Shu et al. (2022) utilizes three sample mining strategies within the context of contrastive learning, aiming to obtain features related to facial expressions. PCL, proposed by Liu et al. (2023), decouples facial expression and pose information, and devises pose-related contrastive learning to extract robust unsupervised facial representations. These methods commonly face an issue when applying contrastive learning, where they construct positive–negative sample pairs by considering augmented versions of the same image as positives and other images as negatives. However, they often overlook the side effects of false-negative pairs belonging to the same category.

2.2 Contrastive Learning

Contrastive Learning (CL) is a self-supervised learning method that learns useful feature representations by training models to bring similar samples closer together and push dissimilar samples away. SimCLR (Chen et al., 2020) proposes to use data augmentation to generate sample pairs and then train the network by maximizing their similarity. MoCo (He et al., 2020) increases the number of negative samples in contrastive learning by maintaining a negative sample queue. Chen and He (2021) proposed the SimSiam method to avoid collapsing solutions by maximizing two data augmentation images of a picture. These methods make use of data augmentation, contrast loss, momentum encoders, and memory banks to enhance representation learning on unlabeled data. Recently, there are methods such as Dwibedi et al. (2021); GE et al. (2023); Shu et al. (2022) that address the issue of treating samples of the same category as negative samples in CL by identifying false negative samples through nearest neighbor analysis.

2.3 Nearest Neighbor Exploration in Visual Recognition

Many computer vision tasks have made wide use of Nearest Neighbor (NN) algorithm, such as image classification (McCann and Lowe, 2012), object detection/segmentation (Harini and Chandrasekar, 2012), and domain adaptation (Yang et al., 2021). It is commonly used to explore the comprehensive relationships between samples, providing convenience for various computer vision tasks.

In the field of self-supervised learning, there are also emerging methods that investigate the use of the NN algorithm. For instance, in CL frameworks, SwAV (Caron et al., 2020) employs a method different from direct feature

comparison by additionally constructing prototype feature clusters to maintain consistency between sample features and representations. NNCLR (Dwibedi et al., 2021) uses an explicit support set to find nearest neighbors. SNCLR (GE et al., 2023) uses an attention module to get the correlation between the neighbors and the current sample. Overall, these methods have achieved certain success in using cosine similarity to determine neighbor positions. However, there still exists a considerable gap between the data similarity at that time and the true similarity obtained from labels. This can potentially affect the learning of nearest neighbor relationships during the training process, thereby influencing the quality and performance of representation learning. Therefore, accurately estimating the similarity between data samples remains a promising direction for further enhancing the effectiveness of self-supervised learning methods.

3 Method

3.1 Overview

The overview of our proposed Pose-disentangled Contrastive Facial Representation Learning (PCFRL) framework is illustrated in Fig. 2. We devise a three-component framework for PCFRL. The first component disentangles pose-aware features from non-pose face-aware features based on our previous work, PCL (Liu et al., 2023). After feature disentanglement, we observe that the same person or pose might generate false negative sample pairs for CL, which may introduce significant confusion. To address this, we introduce a false negative pair module as the second component to further improve the learning procedure based on a novel neighborhood-cohesive pair alignment (NPA) method. Lastly, the third component includes two calibrated CL losses, namely pose-aware calibrated CL loss and face-aware calibrated CL loss, for learning with pose-aware features and non-pose face-aware features. The calibrated CL losses adaptively optimize the model with calibrated false negative pairs based on the NPA results, further reducing the risk of confusion that would affect learning procedures. In the following sections, we first briefly review the PCL method (Liu et al., 2023), and then introduce the details of the other components in our PCFRL framework, including false-negative pair calibration and modified CL losses.

3.2 Revisiting PCL

In our original study, PCL (Liu et al., 2023), we predominantly seek to develop an effective feature disentanglement mechanism for improving the CL on both pose-related information and pose-unrelated face information. To achieve this, we introduced a pose decoupling decoder (PDD) for the pose disentanglement.

Specifically, the PDD identifies and separates pose-related face features from pose-unrelated face features based on representation reconstruction. The general concept is to make sure that the same face image with a different pose can be reconstructed based on the new pose feature and the original non-pose face feature. As a result, the pose feature would become more sensitive to pose changes and the non-pose face feature would be consistent after pose alteration, thereby satisfying the disentanglement purpose.

Mathematically, we denote a face image as s with a pose p. We then transform p using some augmentation techniques like flipping, obtaining the pose-augmented face image $\hat{s}$ with the new pose $\hat{p}$. By taking s as input, the PDD will tend to extract a pose-aware feature $\vec{F}_p$ and a non-pose face-aware feature $\vec{F}_f$. Similarly, if taking $\hat{s}$ as input, we can have extracted features $\vec{F}_{\hat{s}}$ and $\vec{F}_{\hat{p}}$, respectively. Subsequently, we introduce a series of reconstruction objectives, forming:

$$\begin{aligned} L_{dis} = &||s - D(\vec{F}_f, \vec{F}_p)||_1 + ||\hat{s} - D(\vec{F}_f, \vec{F}_{\hat{p}})||_1 \\ &+ ||s - D(\vec{F}_{\hat{f}}, \vec{F}_p)||_1 + ||\hat{s} - D(\vec{F}_{\hat{f}}, \vec{F}_{\hat{p}})||_1, \end{aligned} \tag{1}$$

where $||\cdot||_1$ represents l_1-norm, D is the additional reconstruction network that is used to translate the extracted two types of features into the reconstructed face.

In addition to the above-mentioned reconstruction loss, we further introduce an orthogonal loss to make the disentangled pose-aware feature and non-pose face-aware feature orthogonal to each other, minimizing the possibility that both types of features contain redundant information. This loss is written as:

$$L_{orth} = \frac{1}{N}\left(\sum_{i=1}^{N} ||\vec{F}_f \cdot \vec{F}_p||_2^2 + \sum_{i=1}^{N} ||\vec{F}_{\hat{f}} \cdot \vec{F}_{\hat{p}}||_2^2\right), \tag{2}$$

where N is the number of samples in a training batch.

With loss functions defined in Eqs. 1 and 2, we can train the PDD using the sum of the two losses: $L_{PDD} = L_{dis} + L_{orth}$. The PDD can be optimized to disentangle pose-aware features from non-pose face-aware features effectively, which further improves contrastive learning-based SFRL performance. Our original study proves that this mechanism already achieves promising improvement on several downstream tasks. For more details, we refer readers to the Liu et al. (2023).

3.3 False-negative Pair Calibration for Disentangled Features

Following the disentanglement of pose-aware features and non-pose face-aware features, our original PCL work directly

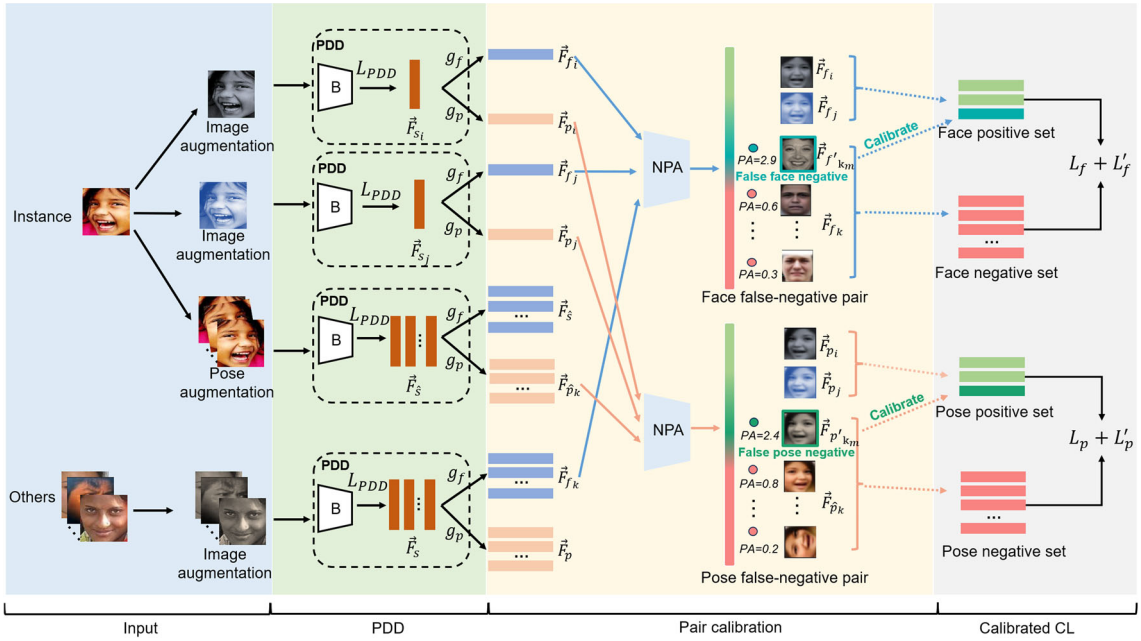


Fig. 2 Detailed pipeline of our proposed PCFRL for pose-aware self-supervised facial representation learning. Building upon the PDD, we first disentangle the pose-aware features from non-pose face-aware features. Then, we further introduce the false-negative pair calibration module to calculate the neighbor-cohesive pair alignment (NPA) scores, resulting in the calibrated pose-aware and face-aware false-negative pairs, respectively. Moreover, with the calibrated false-negative pairs, we devise two calibrated CL losses, namely calibrated pose-aware CL and face-aware CL, to facilitate the development of more robust pose-aware facial representation

applies contrastive learning (CL) to both types of features. In a typical CL procedure, sample selection is used to generate positive and negative pairs for learning. In particular, a pair of samples from the augmentation of the same image is usually considered a positive pair, while a pair of samples from different images is a negative pair. However, drawing negative pairs from different images would be unsatisfactory. Specifically, the selected positive and negative pairs for CL should be different for learning with pose-aware features and non-pose face-aware features. Intuitively, the pose-aware features should favor the pairs of samples with the same pose as positive pairs and others as negative pairs, while the non-pose face-aware features should consider the pairs of samples with the same facial characteristics as positive pairs and others as negative pairs. If we only use shared positive and negative pairs, the CL on both types of features could be compromised due to potentially significant confusion. For example, the two images from different persons with a shared pose would be selected as a negative pair which is inappropriate for CL on pose-aware features. It is worth noting that the pairs that should be calibrated only come from the negative pairs. This is because we can correctly generate positive pairs based on augmentation operations, while it is difficult to ensure correct negative pairs when using different images in a batch. To avoid this, we need to identify incorrect negative pairs and calibrate them into positive pairs.

In general, to identify and calibrate these false-negative pairs, we propose a neighborhood-cohesive pair alignment (NPA) method. The NPA method first estimates the neighborhood-cohesive pair alignment scores for negative pairs obtained using a typical sample selection procedure. Then, using the pair alignment scores, we identify false-negative pairs. Subsequently, we introduce a thresholding-based false-negative pair calibration algorithm to calibrate the identified false-negative pairs into positive pairs.

False-negative Pairs According to the above discussion, we formulate false-negative pairs as inappropriate negative pairs for CL regarding either pose-aware features or non-pose face-aware features.

Neighborhood-cohesive Pair Alignment Score Our false-negative pair calibration mechanism is mainly based on the estimation of an alignment score between a pair of samples. Specifically, using the positive/negative pairs obtained in typical CL procedures and our original PCL paper, we tend to align two samples in a negative pair. If we find that the two samples in a negative pair align with each other, we then tend to consider this pair as a false-negative pair which can be calibrated to a positive pair. To achieve sample alignment, we introduce the NPA procedure.

The introduction of NPA is inspired by the near-neighbor relation learning in graph model (Yu et al., 2021), which validates that if samples "A" and "B" are both consistent with

sample "C", samples "A" and "B" can also be considered as consistent. In our framework, we use both cosine similarity and the neighborhood samples (i.e., other samples in the same batch) to generate an alignment score between sample "A" and sample "B" in a negative pair. Then, a high alignment score can suggest that "A" and "B" in a negative pair are likely to contain similar information. Mathematically, our NPA calculates the alignment score according to:

$$PA\left(v_i, v_j\right) = cos\left(v_i, v_j\right) + \alpha NS\left(v_i, v_j\right), \tag{3}$$

where PA denotes the pair alignment score, cos represents a cosine similarity, NS represents a neighborhood-cohesive sample consistency score, and α is a trade-off parameter that determines the importance of NS over cos.

The cosine similarity between a pair of samples, e.g., the i-th and j-th sample, can be calculated based on their corresponding feature vectors v_i and v_j:

$$cos\left(v_i, v_j\right) = \frac{v_i^T v_j}{\|v_i\|_2 \|v_j\|_2}, \tag{4}$$

where $\|\|_2$ is the L2-norm of vectors, and v_i^T is the transpose of the v_i.

We formulate the neighborhood-cohesive sample consistency estimation as the following procedure:

$$NS(v_i, v_j) = \sum_{k \neq i,j}^{2N} cos\left(v_i, v_k\right) cos\left(v_k, v_j\right), \tag{5}$$

where $2N$ is the total number of faces and their augmented samples in a training batch, and i, j, k index the samples in the training batch. The calculation procedure of NPA is briefly shown in Fig. 3.

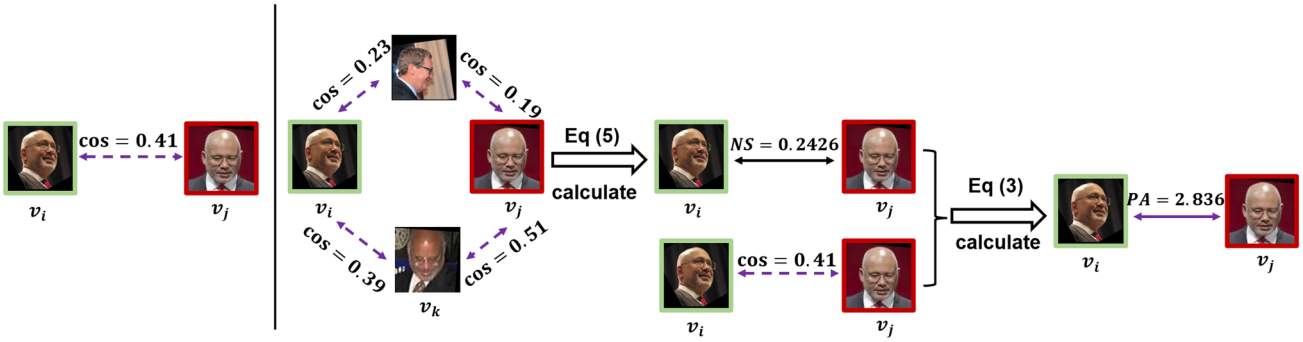
Using Eq. 3, we can calculate the pair alignment score appropriately. Considering that we disentangled pose-aware features $\vec{F}_p$ from non-pose face-aware features $\vec{F}_f$, we use $\vec{F}_{p_i}, \vec{F}_{p_j}$ to represent the pose-aware features of i-th and j-th samples in the same batch, respectively. Then, by substituting v_i and v_j in Eq. 3 with $\vec{F}_{p_i}$ and $\vec{F}_{p_j}$, we obtain the pair alignment between the pose-aware features extracted from the i-th and j-th samples. Similarly, the pair alignment between non-pose face-aware features of i-th and j-th samples can be estimated by substituting v_i and v_j in Eq. 3 with $\vec{F}_{f_i}$ and $\vec{F}_{f_j}$, respectively.

Discussion 1: Relation to Common Similarity Estimation. According to Eq. 3, our pair alignment score uses a cosine similarity and a neighborhood-cohesive sample consistency score, as shown in Fig. 3. We would like to mention that this is not the same as a common similarity estimation procedure. Although common similarity scores like cosine similarity can provide a promising estimation of whether two samples have similar information, these similarity calculations might be ambiguous or not sufficiently discriminative for the disentangled features. In particular, using the disentangled high-level feature vectors may make cosine similarity difficult to depict intricate and subtle differences in factors like facial attributions, illumination, and so on. In high-dimensional spaces, many vectors that are supposed to be different might appear similar. For example, in a true-negative pair with two images of different persons, the two images may have the same lighting conditions, and the cosine similarity might generate a considerably high similarity score that would lead to inappropriate calibration. As a result, despite many existing methods (Dwibedi et al., 2021; GE et al., 2023) that primarily rely on cosine similarity (Rahutomo et al., 2012) to estimate similarities between two samples, we propose that comparing with the neighborhood samples can provide a more holistic estimation of sample alignment. Neighborhood samples would add an extra layer of discrimination by considering the collective relationships among samples, which could be particularly beneficial for the disentangled features.

Discussion 2: Relation to Near-neighbor Relation Learning. The difference between Eq. 5 and the near-neighbor relation learning in graph model (Yu et al., 2021) is that we do not consider cosine between i-th and j-th sample. We design this for two reasons. Firstly, excluding cosine between i-th and j-th sample enhances the coherence estimation among i-th or j-th sample and other samples, which can alleviate the limitations of cosine between i-th or j-th sample themselves as discussed previously. This could consider the broader context in relation to other samples in the dataset, not just the pairwise relationship. Secondly, as written in Eq. 3, we weightly sum the NS and cosine between the i and j sample, which enables us to adjust the importance of Eq. 5 over Eq. 3 via α.

Thresholding-based False-negative Pair Calibration With the calculation of neighborhood-cohesive pair alignment scores based on Eq. 3, it is then possible to identify false-negative pairs. If the i-th and j-th sample are from different images but has a high alignment score, we then consider them as a false-negative pair and calibrate them to a positive pair.

To perform the calibration, we introduce a thresholding-based procedure. Our detailed procedure is as follows. Firstly, we split the samples of a batch into two groups: Group A contains the samples that will be augmented to generate positive pairs; Group B contains the samples that are different from Group A samples. Typically, when generating negative pairs, a sample i' from Group A and a sample j' from Group B will form a negative pair. To identify false-negative pairs, we first draw a sample i' from Group A and compare it with all the samples from Group B according to our proposed NPA method described in Eq. 3, obtaining a set of pair alignment scores $\{PA(v_{i'}, v_{k'}) | i' \in A\}$. We then find the negative sample k'_m from Group B that has the maximum similarity



(a) cosine similarity **(b) neighborhood-cohesive pair alignment**

Fig. 3 Conceptual comparison between the cosine similarity and our proposed neighborhood-cohesive pair alignment score. **a** Cosine similarity access the face relationship between two samples. **b** Our neighborhood-cohesive pair alignment score measures the face relationship by using all neighbor samples' similarities in a training batch. Provide an example in (**b**) to illustrate. First, according to Eq. 5, calculate the dot product of the cosine similarity between samples i and j and other samples to obtain NS. Then, according to Eq. 3, add the result to the cosine similarity between these two samples to obtain NPA

score with i': $k'_m = \arg max(\{PA(v_{i'}, v_{k'})|k' \in B\})$. If this maximum pair alignment score $PA(v_{i'}, v_{k'_m})$ is larger than a threshold T, we then calibrate (i', k'_m) from the negative pair to both the positive pair and the calibrated pair (see the $\mathcal{P}$ and $\mathcal{C}$ in Algorithm 1). Otherwise, we follow the typical procedure and randomly draw a sample j' from Group B to form a negative pair with i'. We present detailed steps in Algorithm 1. It is worth mentioning that the disentangled pose-aware features and non-pose face-aware features need to calibrate their negative pairs respectively. Regarding this, the Algorithm 1 will run twice, once for pose-aware learning and once for non-pose face-aware learning. Also note that the neighborhood cohesive calculation as described in Eq. 5 will be calculated across all the samples in the batch, containing both Group A and Group B samples.

Discussion 3: Compared to Other Calibration Methods. Other methods like Dwibedi et al. (2021) identify the negative pairs with the top-K neighbor coherence values as false-negative pairs and calibrate these pairs into positive pairs accordingly. However, we found that the existing top-K false negative pair calibration mechanisms are not optimal for the disentangled features in our framework. Firstly, the top-K mechanism does not explicitly account for actual alignment scores and always calibrates K samples, which may also calibrate the true negative pairs that should remain as negative. Secondly, since the K is fixed, this may either not cover all the false negative pairs or cover too many negative pairs. Although top-K methods may be effective for general contrastive learning, it is not appropriate for our disentangled features, since the pose-aware pairs and non-pose face-aware pairs would suffer differently from the false-negative pair problem. Alternatively, our devised thresholding-based false-negative pair calibration mechanism relies on appropriate alignment scores and can be more adaptable to different types of pairs (e.g., pose-aware pairs and non-pose face-aware pairs). We will discuss the difference between the top-K method and our thresholding-based method in the Experiment section.

Algorithm 1. False-negative Pair Calibration.

Input:
A random sample i' from Group A (group of samples to be augmented) in a training batch;
Group B samples (samples not in Group A) in the same training batch;
Current set of negative pairs for CL denoted as $\mathcal{N}$;
Current set of positive pairs for CL denoted as $\mathcal{P}$;
Current set of calibrated false-negative pairs, $\mathcal{C}$.
Output:
Updated set of negative pairs, $\mathcal{N}$;
Updated set of positive pairs, $\mathcal{P}$;
Updated set of calibrated false-negative pairs, $\mathcal{C}$;
Compute neighborhood-cohesive pair alignment scores based on Eq. 3, obtaining $\{PA(v_{i'}, v_{k'})|k' \in B\}$;
$k'_m \leftarrow \arg max\{PA(v_{i'}, v_{k'})|k' \in B\}$;
if $PA(v_{i'}, v_{k'_m}) \geq T$ **then**
  $\mathcal{P} \leftarrow \mathcal{P} \cup (i', k'_m)$;
  $\mathcal{C} \leftarrow \mathcal{C} \cup (i', k'_m)$;
else
  Randomly draw a sample j' from Group B;
  $\mathcal{N} \leftarrow \mathcal{N} \cup (i', k')$
end
return $\mathcal{P}, \mathcal{N}, \mathcal{C}$

3.4 Contrastive Learning with Calibrated Pairs

Although contrastive learning can be directly applied to all the pairs obtained after calibration, we found that the potential differences between samples in the calibrated false-

negative pairs might still introduce confusion. For example, there may be subtle differences for samples in a calibrated false-negative pose pair, which may introduce confusion if considering these two samples as the same pose. Accordingly, we devise two novel calibrated CL losses for learning the calibrated false-negative pose-aware pairs and false-negative face-aware pairs, respectively. In the following, we first introduce the normal CL loss for learning with normal non-calibrated pairs and the new calibrated CL loss for calibrated false-negative pairs.

Before introducing the detailed loss formulation, we follow our Algorithm 1 and use $\mathcal{P}$ to represent the positive set, $\mathcal{N}$ to represent the negative pair set, and $\mathcal{C}$ to represent a set of calibrated pairs. Since the calibrated set $\mathcal{C}$ consists of the pairs calibrated from negative to positive, thus $\mathcal{C} \subset \mathcal{P}$.

Contrastive Loss for Normal Pairs. For normal non-calibrated positive and negative pairs, we follow the typical CL formulation to define the loss function. Then, we have the learning loss for a positive pair $v_i, v_j, ((v_i, v_j) \in \mathcal{P} - \mathcal{C})$:

$$L(v_i, v_j) = -log\left(\frac{\exp(cos(v_i, v_j)/\tau)}{\sum_{(v_i,v_k)\in\mathcal{N}} \exp(cos(v_i, v_k)/\tau)}\right), \quad (6)$$

where temperature parameter τ controls the smoothness of the similarity scores, i, j indexes over $\mathcal{P}$ but NOT including calibrated pairs in $\mathcal{C}$, and k indexes over $\mathcal{N}$. Here, the v_i and v_j represent the feature vectors extracted from the related pair of samples.

In our case, v_i and v_j can be substituted by both the pose-aware features $\vec{F}_{p_i}, \vec{F}_{p_j}$ and the non-pose face-aware features $\vec{F}_{f_i}, \vec{F}_{f_j}$. If calculated over pose-aware features, we obtain pose-aware contrastive loss $L_p = L(\vec{F}_{p_i}, \vec{F}_{p_j})$. Similarly, we have a non-pose face-aware contrastive loss $L_f = L(\vec{F}_{f_i}, \vec{F}_{f_j})$. It is worth noting that we introduce two different data augmentation methods, i.e., pose augmentation and image augmentation, for constructing pose-aware and non-pose face-aware positive pairs, respectively. Details about these augmentations can be seen Sec. 4.1.

Calibrated Contrastive Learning Losses For the calibrated pairs in $\mathcal{C}$, we devise new calibrated contrastive learning losses for pose-aware features and non-pose face-aware features, respectively. In order to reduce the risk of confusion, we introduce the calibrated CL losses with an adaptive weighting similarity for the calibrated pairs, i.e., the positive loss weighted based on the neighborhood-cohesive pair alignment score and cosine similarity. If two samples in a calibrated pair have a high alignment score, we then increase their importance in CL, otherwise, we decrease their importance weights, resulting in a more adaptive optimization scheme compared to the conventional CL scheme that treats all positive pairs equally. That is, for a calibrated pair $(v_{i'}, v_{k'_m}) \in \mathcal{C}$, we re-write the Eq. 6 as follows:

$$L'(v_{i'}, v_{k'_m}) = -log\left(\frac{w_{i'} \cdot \exp(cos(v_{i'}, v_{k'_m})/\tau))}{\sum_{(v_{i'},v_{k'})\in\mathcal{N}} \exp(cos(v_{i'}, v_{k'})/\tau)}\right), \quad (7)$$

where $w_{i'} = \beta \cdot PA(v_{i'}, v_{k'_m})$ assigns weights based on the pair alignment score for $v_{i'}, v_{k'_m}$, β controls the strengths of the weighting, $\beta \in [0, 1]$, and L' represents the calibrated contrastive loss for the calibrated pair $(v_{i'}, v_{k'_m})$.

In particular, by substituting $v_{i'}$, $v_{k'_m}$ with pose-aware features $\vec{F}_{p'_i}$, $\vec{F}_{p'_{k_m}}$ and the non-pose face-aware features $\vec{F}_{f'_i}$, $\vec{F}_{f'_{k_m}}$ in Eq. 7. If calculated over pose-aware features, we obtain calibrated pose-aware contrastive loss $L'_p = L(\vec{F}_{p'_i}, \vec{F}_{p'_{k_m}})$. Similarly, we have a non-pose calibrated face-aware contrastive loss $L'_f = L(\vec{F}_{f'_i}, \vec{F}_{f'_{k_m}})$. It is worth mentioning that, despite using similar substitutions, pose-aware and non-pose face-aware learning objectives do not share the same positive and negative training pairs. This is because the two learning cases may have different calibrated pairs. As a result, in L'_p and L'_f, we used different β for optimization. We discuss the effects of β in the experiments, showing that the best performance is obtained when $\beta = 0.2$ in L'_p for calibrated pose-aware contrastive learning and $\beta = 1$ in L'_f for calibrated face-aware contrastive learning (see Fig. 6b).

Discussion 4: Different similarities for the calibrated pairs. In Eq. 7, unlike normal CL loss only uses cosine similarity on positive/negative pair learning, we introduce an adaptive weighting similarity that integrates our proposed NPA score with cosine similarity for the calibrated pair learning. This aims to suppress the the risk of potential inappropriate false-negative pair confusion, resulting in more robust contrastive learning on the calibrated pairs. We experimented with alternative similarity calculations for learning from the calibrated pairs, such as using only NPA score or only cosine similarity, but they yielded unsatisfactory results. The more discussion can be seen Table 8 in the experimental part. We believe that the combination of the two is more adaptive to optimise the correction of the calibrated pairs, i.e., the false-negative pairs.

3.5 Overall Learning Objectives

With the loss function defined in Eq. 7, we can achieve more appropriate calibrated contrsative learning on the calibrated samples. Together with Eq. 6 on non-calibrated normal samples, we can fulfill the SFRL by applying the two types of contrastive losses on both the disentangled pose-aware features and non-pose face-aware features.

To sum up, with the obtained normal contrastive losses L_p, L_f and calibrated contrastive losses L'_p, L'_f, overall learning objectives of our PCFRL framework can be the sum

of the objectives discussed above. Meanwhile, in addition to contrastive learning loss, we follow our original PCL work (Liu et al., 2023) and apply a disentanglement loss L_{PDD} to make the network learn to disentangle pose-aware features and non-pose face-aware features appropriately. The L_{PDD} is actually the combination of Eqs. 1 and 2. Please refer to Liu et al. (2023) for more details. As a result, our complete overall learning objectives can be written as:

$$L = L_{PDD} + \alpha_{pose} \cdot (L_p + L'_p) + \alpha_{face} \cdot (L_f + L'_f) \quad (8)$$

where α_{pose} and α_{face} are two parameters weighting the importance of pose-aware contrastive learning and non-pose face-aware contrastive learning. To avoid hyperparameter tuning, we define the two parameters as dynamic weights whose values are determined by the Dynamic Weight Averaging (DWA) (Liu et al., 2019) during training.

4 Experiments

4.1 Implementation Details

Following contrastive learning, we perform augmentations to help obtain positive training pairs (i.e., augmentation on Group A samples as mentioned in Sec. 3.3). However, we would like to mention that the augmentation methods used to generate positive pair for pose-aware learning and non-pose face-aware learning are different. Specifically, the augmentations, including flipping and rotating, help generate positive pairs for pose-aware learning. The other augmentations, such as Gaussian blur, color jitter, random cropping, and Sobel filtering, are primarily used to help generate positive pairs for non-pose face-aware learning.

Our PCFRL framework is implemented using the PyTorch platform. We trained each model for 1000 epochs using the Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and used cosine annealing to reduce the reduction learning rate (0.0001). The batch size and temperature parameter τ in Eq. 6 are set to 256 and 0.07, respectively.

Figure 4 shows the network architecture of our backbone network B (see Fig. 2) and its two branch networks. Referring to FaceCycle (Chang et al., 2021), we adopted a shallow network consisting of 10 convolutional blocks, 2 channel attention blocks, and 2 residual basic blocks. The inspiration for the channel attention module comes from self-attention, which we only use to calculate the relationships between channels, not spatial pixels. The subnet consists of 4 layers, namely two 3 × 3 convolutional layers and 2 leakyReLU as activation functions. For more detailed descriptions of the framework, please refer to the PCL (Liu et al., 2023).

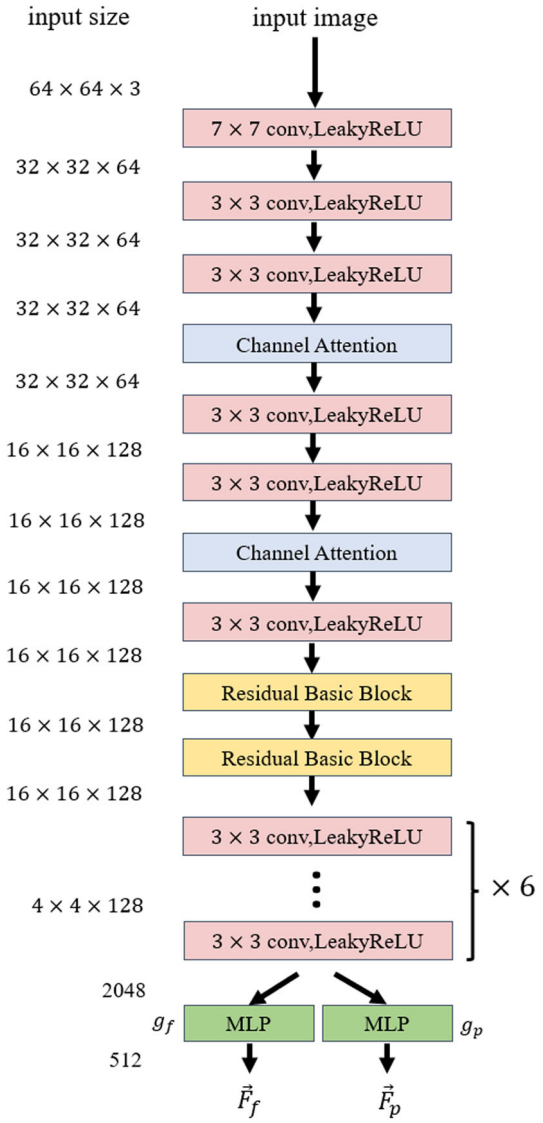


Fig. 4 The detailed network architecture of the backbone and the corresponding two subnets

4.2 Experimental Settings

4.2.1 Datasets

For self-supervised pre-training, we combined two large-scale and in-the-wild face-related datasets, namely the VoxCeleb1 (Nagrani et al., 2017) and VoxCeleb2 (Ojala et al., 2002). The VoxCeleb1 encompasses facial data from 1,251 individuals of diverse age groups and ethnicities. The VoxCeleb2 comprises data from 5,994 individuals. These two datasets together consist of 299,085 video data. We extracted video frames at a rate of 6 frames per second (fps). During training, we resized the extracted video frames to 64 × 64 pixels (Chang et al., 2021).

For facial expression recognition (FER) evaluation, we used two widely-used FER datasets: FER-2013 (Goodfellow et al., 2013) and RAF-DB (Li et al., 2020). FER-2013 contains 28,709 training images and 3,589 testing images. For RAF-DB, we focus on the basic emotion subset, i.e., 12,271 images from RAF-DB are used as training images and 3,068 images are used as testing images.

For face recognition (FR) evaluation, we employed two FR datasets, namely LFW (Wiles et al., 2018) and CPLFW (Zheng and Deng, 2018). LFW comprises 13,233 face images from 5,749 individuals. For the purpose of the FR task, LFW extracts 6000 face pairs, with 3000 pairs consisting of two face images from the same person. CPLFW, on the other hand, includes 3000 pairs of frontal faces captured in different poses to introduce pose variations and thus enhance intra-class differences. The experimental results are averaged over 10 folds.

For facial AU detection evaluation, we used the DISFA dataset (Mavadati et al., 2013) for evaluating our method. The DISFA dataset has a total of 130,788 frames from 26 participants, each labeled with action units of varying intensity levels (from 0 to 5). Frames with intensities greater than 1 are classified as positive, while those with intensities less than or equal to 1 are designated as negative. The experimental results were obtained using a 3-fold cross-validation with reference to Li et al. (2019).

For head pose estimation (HPE) evaluation, we assessed both head pose regression and head pose classification tasks. In the pose regression task, we conducted training using the 300W-LP dataset (Sagonas et al., 2013), comprising 122,450 images, and subsequently performed evaluation on AFLW2000 (Zhu et al., 2016), which includes 2000 images. In the pose classification task, we followed the experimental setup outlined in Liu et al. (2021), and evaluated on the BU-3DFE dataset (Yin et al., 2006), which contains 14,112 training images and 6264 validation images.

4.2.2 Evaluation Protocols

We adopted the standard linear evaluation protocol in SSL-based methods (Chen et al., 2020; Chen and He, 2021; Chen et al., 2021; Chang et al., 2021; Datta et al., 2018; He et al., 2020; Li et al., 2019) for the validation of our method. The linear classifier is a basic fully-connected layer, and it is trained for 300 epochs using the fixed self-supervised face-aware representation $\vec{F}_s$ obtained from the backbone network (see Fig. 2).

For a fair comparison, in line with references such as Chang et al. (2021); Datta et al. (2018); Li et al. (2019), we trained the images with different dimensions in different downstream tasks. Specifically, for the FER task, we resized the images to 100 × 100. The FR task used an image size of 128 × 128, while the HPE applications all employed an image size of 256 × 256.

Table 1 Accuracy comparison of FER performance on FER-2013 and RAF-DB datasets

Method	FER-2013 Accuracy(%)	RAF-DB Accuracy(%)
Fully supervised		
FSN (Zhao et al., 2018)	67.60	81.10
ALT (Florea et al., 2019)	69.85	84.50
Self-supervised (linear evaluation)		
LBP (Ojala et al., 2002)	37.89	52.17
HoG (Dalal and Triggs, 2005)	45.47	63.53
FAb-Net (Jakab et al., 2018)	46.98	66.72
TCAE (Li et al., 2019)	45.05	65.32
BMVC'20 (Lu et al., 2020)	47.61	58.86
MoCo (He et al., 2020)	47.24	68.32
FaceCycle (Chang et al., 2021)	48.76	71.01
SimCLR (Chen et al., 2020)*	49.51	71.06
PCL (Liu et al., 2023)*	56.81	74.47
Ours	**57.30**	**75.68**

Bold numbers represents best results
Note: * indicates that the result is reproduced by authors

4.3 Overall Performance on Different Downstream Tasks

4.3.1 Results on FER

We evaluated the performance on the FER task using the model trained via PCFRL and the state-of-the-art methods. The experimental results are shown in Table 1, demonstrating that our proposed method has better performance compared to other methods. Compared with PCL (Liu et al., 2023), PCFRL achieves a relative improvement of 0.86% on the FER-2013 dataset and an accuracy increase of 1.62% on the RAF-DB dataset. This shows that PCFRL is capable of learning superior self-supervised facial representations.

4.3.2 Evaluation for FR

We also evaluated our method on the face recognition task. The results, as shown in Table 2, indicate that our learned self-supervised pose-aware facial features outperform other methods. Our approach achieved the highest accuracies in both the LFW and CPLFW datasets, at 79.89% and 66.17%, respectively. Compared with the state-of-the-art PCL (Liu et al., 2023), our method achieved a relate increase of 0.21% on the LFW dataset and 2.41% on the CPLFW dataset, indicating that the proposed negative-false pair calibration method and calibrated CL losses effectively boost the performance

Table 2 Accuracy comparison of FR on the LFW and CPLFW datasets

Method	LFW Accuracy(%)	CPLFW Accuracy(%)
Fully supervised		
VGG-Face (Parkhi et al., 2015)	98.95	84.00
SphereFace (Liu et al., 2017)	99.42	81.40
ArcFace (Deng et al., 2019)	99.53	92.08
Self-supervised (Linear evaluation)		
LBP (Ojala et al., 2002)	72.44	-
VGG (Datta et al., 2018)	72.20	-
MoCo (He et al., 2020)*	65.88	57.82
SimCLR (Chen et al., 2020)*	75.97	62.25
FaceCycle (Chang et al., 2021)*	74.12	63.35
PCL (Liu et al., 2023)*	79.72	64.61
Ours	**79.89**	**66.17**

Bold numbers represents best results
Note: * indicates that the result is reproduced by authors

of pose awareness of self-supervised facial representation learning.

4.3.3 Results on Facial AU Detection

PCFRL follows the approach described in Li et al. (2019), which involves employing a binary cross-entropy loss for training a linear classifier in the facial AU detection task. Table 3 reports the experimental results on the DISFA dataset. We compared our method not only with state-of-the-art self-supervised methods, but also with full supervised methods. As can be seen from the table, our PCFRL outperforms the other methods in terms of the average F1 score. Specifically, PCFRL improves the average F1 by 3 points over PCL and 1.8 points over fully supervised learning methods. This suggests that PCFRL achieves more effective facial representation, thanks to the incorporation of neighborhood-cohesive pair alignment for enhanced selection of negative and positive pairs, resulting in a more robust learning process.

4.3.4 Results on HPE

We assessed our method in two pose-related HPE tasks, including head pose regression and pose classification. Following the experimental settings in Liu et al. (2023), head pose regression and classification tasks were pre-trained using the 300W-LP dataset and BU-3DFE, respectively. In the linear evaluation, the head pose regression task was assessed on AFLW2000, while the head pose classification task was evaluated on BU-3DFE. Comparison results with various SSL methods are presented in Table 4. As can be seen from the table, our method surpasses other self-supervised methods in both two tasks, achieving the lowest MAE (12.08%) on AFLW2000 and the highest accuracy (98.96%) on BU-3DFE.

4.4 Performance of Different Similarity Estimation Methods-based Calibration on Contrastive Learning Frameworks

To evaluate the effectiveness of our proposed NPA-based false-negative pair calibration, we applied it to two CL-based self-supervised methods, including PCL (Liu et al., 2023) and SimCLR (Chen et al., 2020). As shown in Table 5, our NPA-based calibration improved the implementation of both SimCLR and PCL in four facial downstream tasks. For instance, in the FER task, our method exhibited relative improvements of 1.62 points over PCL and 0.72 points over SimCLR; in the FR task, it relatively enhanced the accuracy by 2.37 points compared to PCL and 1.93 points compared to SimCLR. These results validate that our method has better facilitation for PCL, which more effectively calibrates both inappropriate pose-aware and face-aware negative pairs, resulting in robust pose awareness facial representation learning.

Moreover, in Table 5, we also compared our NPA method and the cosine similarity (Rahutomo et al., 2012) for false-negative pair calibration in PCL (Liu et al., 2023) and SimCLR (Chen et al., 2020), respectively. The results demonstrate that our proposed NPA method outperforms the cosine similarity method by a significant margin in all four downstream tasks, such as an improvement of 2.27 points for PCL on RAF-DB. This indicates that our proposed NPA-based calibration method can be considered a versatile, plug-and-play module for improving the performance of various CL-related frameworks.

Table 3 Evaluation for Facial AU detection on the DISFA dataset using the $F1$ score

	Methods/AU	1	2	4	6	9	12	25	26	ave
Supervised	DRML (Zhao et al., 2016)	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
	EAC-Net (Li et al., 2017)	41.5	26.4	66.4	50.7	80.5	89.3	88.9	15.6	48.5
	JAA-Net (Shao et al., 2018)	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
Self-superivised	SplitBrain (Zhang et al., 2017)	13.1	10.6	35.7	40.2	30.2	57.5	77.4	40.3	38.1
	DeformAE (Shu et al., 2018)	17.6	12.3	46.7	43.5	26.0	62.7	64.8	**47.6**	40.1
	FAb-Net (Jakab et al., 2018)	15.5	16.2	43.2	**50.4**	23.2	69.6	72.4	42.4	41.6
	TCAE (Li et al., 2019)	15.1	16.2	50.5	48.7	23.3	72.1	72.4	42.4	45.0
	TCAE (Li et al., 2019) *	10.5	13.3	20.9	18.8	7.5	44.7	57.8	9.9	22.9
	FaceCycle (Chang et al., 2021)*	26.4	10.2	37.3	21.5	25.0	71.8	84.2	34.7	38.9
	SimCLR (Chen et al., 2020)*	40.5	46.9	53.8	33.5	24.9	**74.7**	85.0	35.6	49.4
	PCL (Liu et al., 2023)*	53.8	44.9	58.1	37.2	**53.2**	73.1	**86.5**	31.3	54.8
	Ours	**54.5**	**62.1**	**60.3**	36.6	47.4	73.6	86.0	32.6	**57.8**

Bold numbers represents best results
Note: * indicates that the result is reproduced by authors

Table 4 Evaluation on two HPE tasks, including pose regression and classification

	AFLW2000 (pretrained on 300W-LP)				BU-3DFE
	Yaw↓	Pitch↓	Roll↓	MAE↓	Accuracy (%)↑
FaceCycle (Chang et al., 2021)	11.70	**12.76**	12.94	12.47	98.82
MoCo (He et al., 2020)	28.49	16.29	15.55	20.11	75.33
SimCLR (Chen et al., 2020)	11.20	19.86	12.08	14.38	98.85
PCL (Liu et al., 2023)	9.86	16.59	10.62	12.36	98.95
Ours	**9.42**	16.50	**10.32**	**12.08**	**98.96**

Bold numbers represents best results
Note: ↓ represents the smaller is better. ↑ represents the larger is better

Table 5 Performance of NPA-based and cosine similarity-based false-negative pair calibration across various contrastive learning frameworks for four face-related downstream tasks

CL framework	Cosine-based calibration	NPA-based calibration	RAF-DB	CPLFW	DISFA	BU-3DFE
SimCLR (Chen et al., 2020)	x	x	71.06	62.25	49.40	98.85
	✓	x	71.55	63.13	51.78	98.87
	x	✓	**71.57**(**+0.51**)	**63.45**(**+1.2**)	**52.07**(**+2.67**)	**98.88**(**+0.03**)
PCL (Liu et al., 2023)	x	x	74.47	64.61	54.8	98.95
	✓	x	73.41	65.79	50.49	98.84
	x	✓	**75.68**(**+1.21**)	**66.17**(**+1.56**)	**57.80**(**+3**)	**98.96**(**+0.01**)

Bold numbers represents best results

4.5 Ablation Study

4.5.1 Effect of Different Components

In order to evaluate the validity of the main components in our PCFRL, Table 6 presents the ablation study results of gradually introducing the PCL loss (i.e., $L_{PDD} + L_p + L_f$) (Liu et al., 2023), NPA-based false-negative pair calibration (i.e., NPA-calibration), calibrated pose-aware and face-aware contrastive learning loss (i.e., $L'_p + L'_f$) into the baseline (namely SimCLR (Chen et al., 2020)) for three facial downstream tasks, including FER on the RAF-DB dataset (Li et al., 2020), FR on the CPLFW dataset (Zheng and Deng, 2018), and AU detection on the DISFA dataset (Mavadati et al., 2013). Specifically, in the FER task, the baseline achieved 71.06% FER accuracy. Then, we introduced the PCL loss,

resulting in what we now refer to as the PCL (Liu et al., 2023), and achieved an average accuracy of 74.47%. Adding the false-negative pair calibration slightly improves the PCL performance by 0.16% on RAF-DB. The possible reason is that the estimated number of false negatives remains small, and the benefits of NPA may not be pronounced without calibrated contrastive losses. However, when integrated with these losses, NPA significantly improves performance on RAF-DB. Additionally, our NPA-calibration method shows consistent benefits, with a 0.54% and 1.2% direct improvement on CPLFW and DISFA, respectively. Overall, the ablation study suggests that both the proposed false-negative pair calibration and calibrated contrastive learning serve as effective methods to enhance robust self-supervised facial representation.

Fig. 5 Visualization of the pair alignment score (PA) in some training batches, where no false-negative samples can be calibrated because the maximum alignment scores are below the threshold T

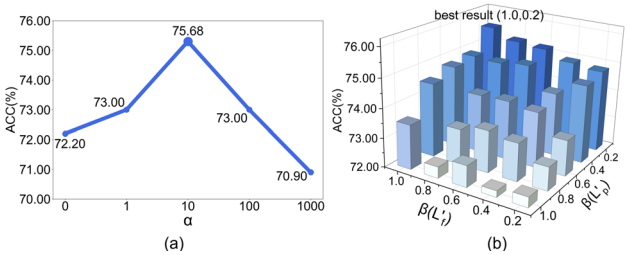


Fig. 6 Effects of key parameters for the FER task on the RAF-DB dataset. **a** Performance with varying α, **b** Performance with varying values of β in two calibrated CL losses

4.5.2 Effects of Different Calibration Methods

We also conducted experiments to analyze the impact of different calibration methods for identifying the false-negative pairs, including the current nearest neighbor method (NN) (Shu et al., 2022), top-K methods (e.g., top-2 and top-5) (Dwibedi et al., 2021; GE et al., 2023), and our thresholding-based calibration method. The comparison results are presented in Table 7. From the results, we can observe that the our proposed thresholding-based calibration method significantly improves the PCL with a relative increase of 1.62 points, indicating that correctly identifying the false negative pairs in a training batch can effectively augment the CL performance. Moreover, we find that calibrating more false-negative pairs, e.g., top-2 and top-5, can affect the CL performance, leading to the significant degradation of performance (about 6.67 points). The reason for this is that when the selected K is greater than 2, the second false negative sample found is likely to be wrong, as shown in Fig. 8. In addition, Fig. 5 illustrates the pair alignment scores (PA) to input instances in some specific training batches. One can see that, in theses batches, the similarity scores are below the established threshold (T=1) in NPA. As a result, it does not generate false negative samples in these training batches. From the figure, our thresholding method successfully avoids labeling these true negative samples, unlike the traditional TOP-K method, which incorrectly labels them as false negatives. This distinction further underscores the reliability of our thresholding method. Instead of directly selecting top-K nearest samples as the potential false-negative pairs, our thresholding-based calibration method obtains the optimal performance by fully considering the distribution of neighborhood-cohesive pair alignment scores in a training batch, with a great robustness and effectiveness.

Table 7 Performance of different sample calibration methods for the FER task on the RAF-DB dataset

Different Strategies	Accuracy on FER (%)
PCL	74.47
Nearest neighbor (Top-1)	75.33
Top-2	71.10
Top-5	67.68
Our thresholding-based calibration	**75.68**

Bold numbers represents best results

4.5.3 Effects of Different Similarity Optimization in Calibrated CL Losses

To thoroughly evaluate the different similarity metrics for the calibrated pairs in calibrated CL losses, we conducted different optimization methods for the calibrated pairs and studied their effects for the FER task on the RAF-DB dataset. Specifically, we implemented the pose-aware and non-pose face-aware calibrated CL losses based on four similarity metrics for calibrated pair optimization: only cosine similarity (Shu et al., 2022), attention-based similarity (GE et al., 2023), and adaptive weighting similarity (ours). The comparison results can be shown in Table 8. One can see that our adaptive weighting similarity obtains the best performance, demonstrating its superior adaptability in optimizing the correction of calibrated pairs.

Table 6 Effects of different modules in our PCFRL on three downstream tasks, i.e., FER, FR, and facial AU detection

Baseline (SimCLR)	PCL	NPA-calibration	$L'_f + L'_p$	RAF-DB	CPLFW	DISFA
✓				71.06	62.25	49.4
✓	✓			74.47	64.61	54.8
✓	✓	✓		74.63	65.15	56.0
✓	✓	✓	✓	**75.68**	**66.17**	**57.8**

Bold numbers represents best results

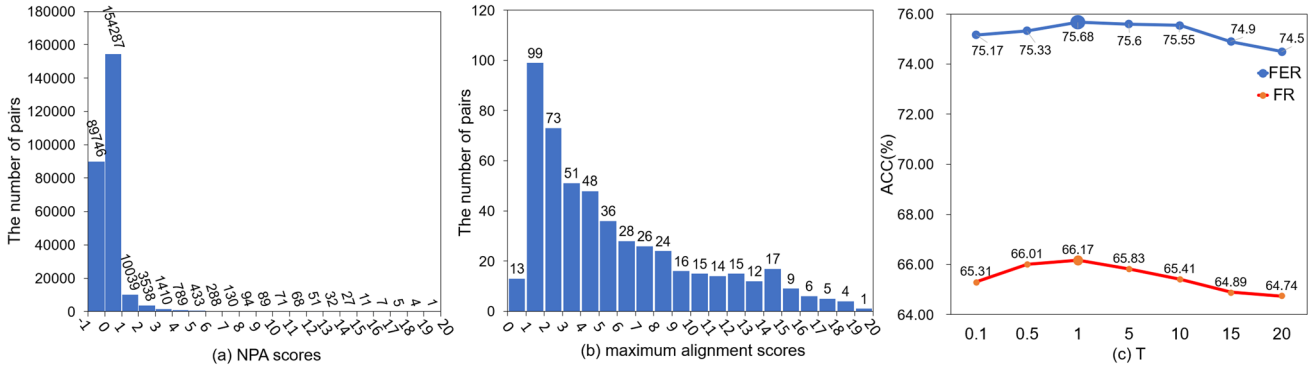


Fig. 7 The distributions of neighborhood-cohesive pair alignment scores, maximum alignment scores, and accuracies. **a** The distribution of the neighborhood-cohesive pair alignment scores for non-pose face-aware features in a training batch, **b** the distribution of maximum alignment score of each sample for face-aware features in a training batch, **c** FER and FR accuracy with various thresholds T

Table 8 Performance of different similarity optimization schemes in calibrated CL losses on the FER task with the RAF-DB dataset

Similarity optimization	Accuracy on FER (%)
Only cosine similarity	73.52
Attention-based similarity	72.98
Only NPA score	73.60
Adaptive weighting similarity (ours)	**75.68**

Bold numbers represents best results

4.5.4 Effects of Key Parameters in PCFRL

We further discuss the influence of the key parameters in our PCFRL approach.

Performance with various α We evaluated the effects of the parameter α in Eq. 3, which is a trade-off parameter that determines the importance of neighborhood-cohesive sample consistency NS over cosine similarity cos. Figure 6a presents the FER accuracy curves with the variation of α. One can see that the accuracy reached the highest 75.68% when we set α to 10. The results show that the neighborhood-cohesive sample consistency NS provides more comprehensive modelling of face sample relationships than cos. Moreover, to verify the reliability of the T setting, Fig. 7c illustrates the impact of varying threshold values (ranging from 0 to 20) on FER and FR performance, respectively. The results show that the trend in threshold selection is consistent across both tasks, with the best performance achieved when $T = 1$ for both FER and FR tasks. Additionally, the slight variations in results across different thresholds indicate that the overall experimental outcomes remain robust across various tasks. By leveraging the distribution of alignment scores within the current training batch, our proposed thresholding-based calibration method demonstrates greater adaptability to both types of sample pairs: pose-aware and non-pose face-aware pairs.

Performance with various β In addition, within our proposed calibrated CL loss, we conducted a detailed evaluation of the effects of its hyperparameters, i.e., β, as defined in Eq. 7. Figure 6b illustrates the performance with varying β in the calibrated pose-aware CL loss L'_f and calibrated face-aware CL loss L'_p, respectively. From the Fig. 6b, we observe that the optimal results are achieved when β is set to 1 for L'_f and 0.2 for L'_p, respectively.

Performance with various T *in calibration* The threshold T in calibration is an adaptive parameter for determining false-negative pairs in the pose-aware features and non-pose face-aware features, respectively, in each training batch. Instead of using the top-K method for calibrating false-negative pairs, we introduce a adaptive threshold T for dynamically calibrating the inappropriate pairs by analyzing the distribution of the maximum alignment score among all negative pairs in each training batch. In particular, Fig. 7a shows the distribution of neighborhood-cohesive pair alignment scores between negative pairs in the training batch,

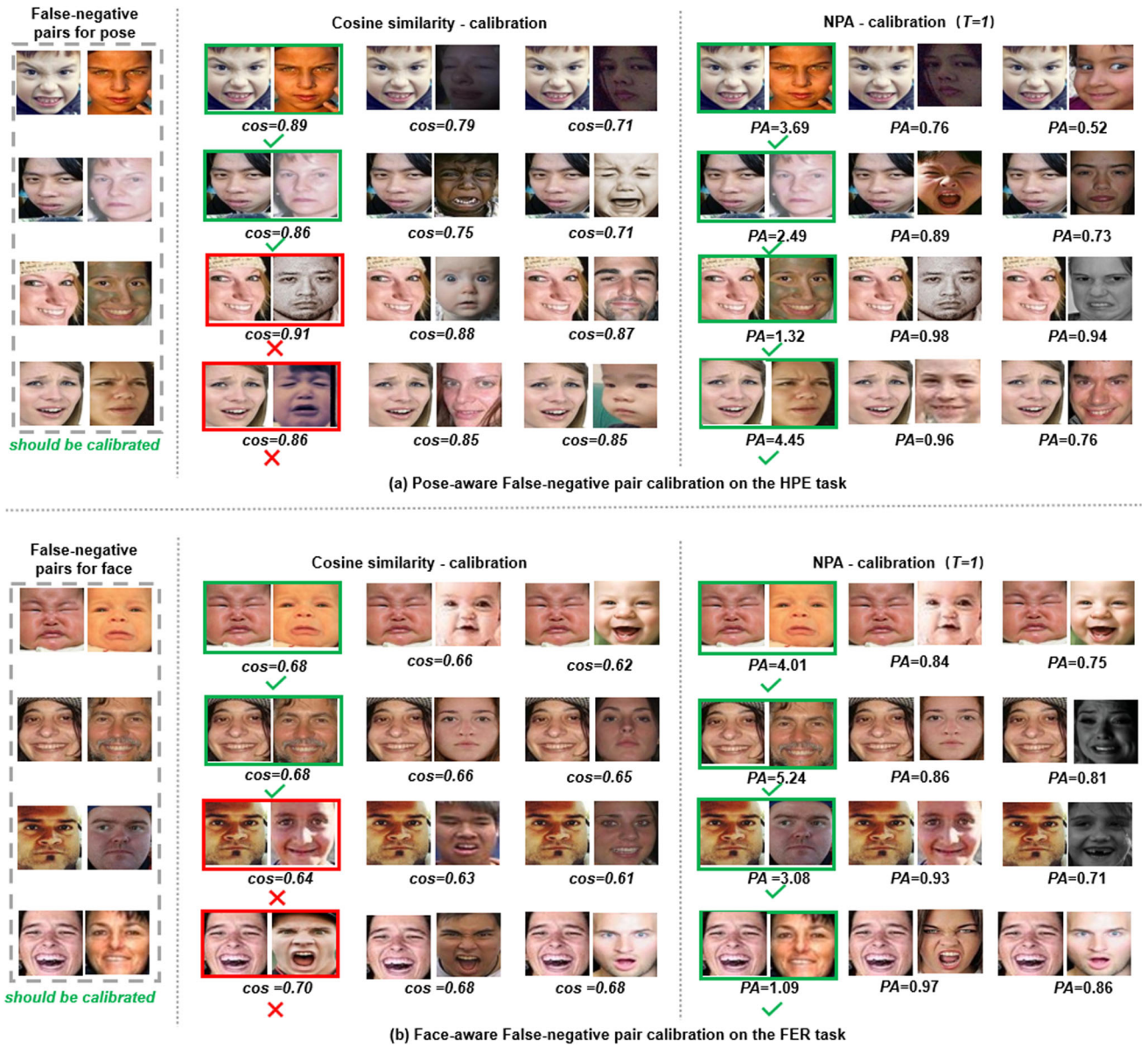


Fig. 8 Comparison of cosine similarity (cos) and our proposed neighborhood-cohesive pair alignment neighborhood-cohesive pair alignment score (PA) for false-negative pair calibration. **a** Calibration for pose-aware false-negative pairs on the HPE task, **b** calibration for non-pose face-aware false-negative pairs on the FER task. It's worth noting that the calibration threshold is set to $T = 1$ in this training batch

and Fig. 7b displays the distribution of the maximum alignment scores for each sample in the training batch. According to the two distributions, we observed that most of both neighborhood-cohesive pair alignment scores and the maximum alignment scores are distributed around to 1. Therefore, we set T to 1 in this training batch. Similarly, in another training batch, we conducted the similar threshold setting process. Moreover, to verify the reliability of the T setting, Fig. 7c illustrates the impact of various threshold values ranging from 0 to 20 on FER performance. From the results, it is evident that our thresholding-based calibration achieves the best performance, i.e., $T = 1$ selected in this batch. By relying on the distribution of alignment scores in the current training batch, our proposed thresholding-based calibration is more adaptable to the two types of sample pairs, namely pose-aware pairs and non-pose face-aware pairs.

Table 9 Accuracy on image classification tasks for STL10 and CIFAR10 datasets

SimCLR	NPA-calibration	SLT10	CIFAR10
✓		75.39	70.32
✓	✓	**76.30**(+**0.91**)	**71.14**(+**0.82**)

Bold numbers represents best results

Table 10 Comparison of computational complexity

	Time (s)	FLOPs (G)	MACs (G)	Params (M)
PCL	0.0547	17.53	7.56	16.8
Ours	0.0562	17.53	7.56	16.8

4.5.5 Generalization for General Image Classification

To further validate the generalization capabilities of our NPA-calibration method, we conducted experimental research on two standard image recognition tasks using the SimCLR model integrated with our proposed approach. The experiments were carried out on two widely used image classification datasets, namely STL10 (Coates et al., 2011) and CIFAR10 (Krizhevsky and Hinton, 2009). The results in Table 9 demonstrated significant performance enhancements, with our method yielding a 0.91% improvement in accuracy on the STL10 dataset and a 0.82% improvement on the CIFAR10 dataset. These findings not only underscore the effectiveness of the NPA-calibration method but also highlight its generalization potential across other self-supervised learning frameworks in diverse visual tasks.

4.5.6 Analysis of Computational Complexity

Here we have selected four indicators to evaluate the computational complexity of our proposed approach, i.e., the time taken to run an epoch, FLOPs, MACs, and params. The experimental results are shown in Table 10. As we can see from the table, the FLOPs, MACs, and params values have not changed, because our method has not changed the backbone of the model. And the time it takes to run an epoch is not significantly increased as you can see from the table. Therefore, compared to PCL, our approach obtains a performance improvement without introducing an increase in computational complexity.

4.6 Qualitative Analysis and Visualization

4.6.1 Visualization on False-negative Pair Calibration for Poses

To illustrate the qualitative results of calibration, Fig. 8a shows the comparison of pose-aware false-negative pair calibration via our NPA method and cosine similarity (Dwibedi et al., 2021). In the figure, the first column shows the false-negative pairs that should be calibrated, followed by showing the pair calibration obtained by cosine similarity and our NPA, respectively. The cosine similarity yielded several inappropriate results (see the red boxes), suggesting that this metric could not accurately identify the false negative pairs from pose-aware features. In contrast, our NPA correctly identified all false-negative pairs that have the same poses due to the comprehensive consideration of the relationships among all coherent neighboring samples.

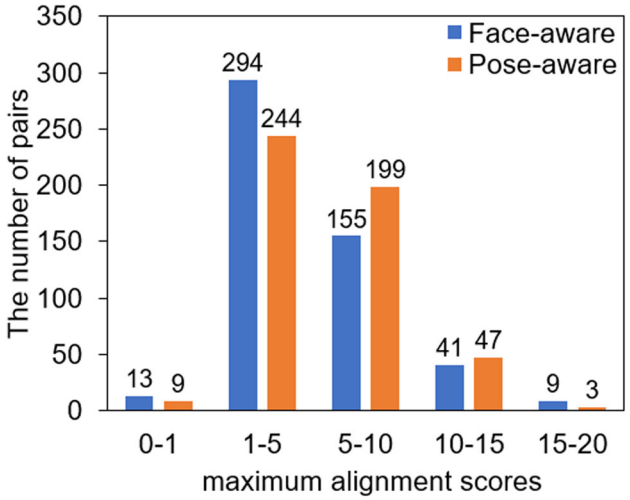


Fig. 9 The distributions of the maximum alignment scores for pose-aware and face-aware negative pairs. The distribution discrepancy indicates that the pose-aware false-negative pair calibration is significantly different from the face-aware false-negative pair calibration

4.6.2 Visualization on False-negative Pair Calibration for Non-pose Faces

Figure 8b presents a comparison of false-negative pair calibration from non-pose face-aware features by using our NPA and the cosine similarity (Dwibedi et al., 2021) on the FER task. One can see that our NPA method effectively calibrates all face-aware false-negative pairs that have the same facial expressions. It provides further evidence that our NPA method outperforms the conventional cosine similarity-based calibration, thanks to comprehensive consideration of relationships among all coherent neighborhood samples.

4.6.3 Calibration Discrepancy between Pose-aware and Face-aware Pairs

In Fig. 9, we display the distributions of the maximum alignment scores for pose-aware and face-aware pairs in a training batch. According to the distributions, it is evident that the calibrated pose-aware and face-aware false-negative pairs

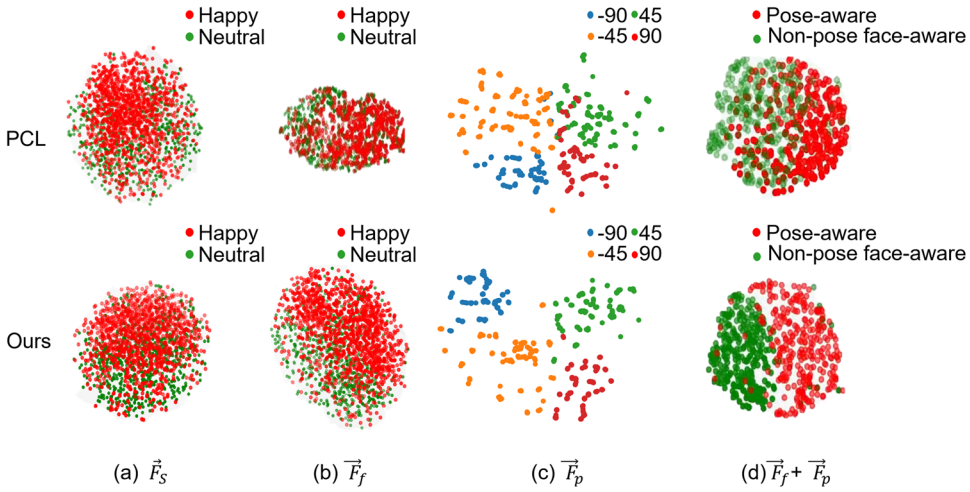


Fig. 10 The features learned by PCL and our PCFRL respectively in t-SNE feature visualization. **a** Self-supervised facial features extracted from PCL and our PCFRL for FER (on RAF-DB), respectively, **b** non-pose face-aware features extracted from PCL and our PCFRL for FER (on RAF-DB), respectively, **c** non-face pose-aware features extracted from PCL and our PCFRL for HPE (on BU-3DFE), respectively, **d** the disentangled pose-aware and non-pose face-aware features via PCL and PCFRL, respectively

through NPA exhibit significant differences. This also shows that our proposed thresholding-based calibration procedure is specifically important because the pose-aware features and non-pose face-aware features do not share the same positive and negative-pair selection.

4.6.4 Feature Visualization

In Fig. 10, we utilized t-SNE to visualize the feature distributions obtained by PCL and PCFRL, respectively. To facilitate a clear comparison, we visualized (a) the self-supervised facial features $\vec{F}_s$ before the PDD module for the FER task, (b) non-pose face-aware features $\vec{F}_f$ for the FER task, (c) pose-aware features $\vec{F}_p$ for the HEP task, and (d) the disentangled features $\vec{F}_p + \vec{F}_f$. Comparing these features learned by the previous PCL (Liu et al., 2023), our improved PCFRL obtained more discriminated self-supervised representations. This is attributed to the more accurate false-negative pair calibration, indicating a stronger facial representation learning capability compared to PCL.

5 Conclusion

In this paper, we propose a novel Pose-disentangled Contrastive Facial Representation Learning framework, called PCFRL, for pose awareness self-supervised facial representation. To achieve this, PCFRL introduces a novel neighborhood-cohesive pair alignment method to calibrate false-negative pairs with pose-aware features and non-pose face-aware features. Moreover, two new calibrated CL losses are devised to dynamically learning on the calibrated pairs via an adaptive weighting strategy, ultimately enhancing the learning of robust, pose-aware self-supervised facial representations. The effectiveness of our proposed PCFRL is demonstrated in four face-related downstream tasks, including FER, FR, facial AU detection, and HPE. Extensive experiments show PCFRL's superiority in enhancing the learning capacity compared to the previous PCL method.

Despite the significant achievements, there are still areas in which our PCFRL could be improved. For instance, intrinsic noises within facial images, such as variations in illumination, shadows, occlusions, and so on, can be challenging to disentangle, thereby affecting PCFRL's performance. In the future, we plan to introduce physics-informed prior knowledge to further disentangle these complex noises for robust unsupervised facial representation.

Acknowledgements This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

Author Contributions Yuanyuan Liu, Shaozhe Feng, and Zhe Chen contributed the central idea, analyzed most of the data, and wrote the first draft of the paper. The remaining authors dedicated themselves to refining these ideas, conducting additional analyses, and ultimately completing the paper.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data Set Availability All datasets utilized in our research are publicly available and licensed for use. To ensure fairness in experimental comparisons, the data partitioning and evaluation of our experiments followed the related work to maintain consistency. Download links for the relevant datasets can be found at https://github.com/fulaoze/CV/tree/main.

Declarations

Conflict of interest There is no conflict of interest in our work.

References

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems, 33*, 9912–9924.

Chang, J.-R., Chen, Y., & Chiu, W.-C. (2021). Learning facial representations from the cycle-consistency of face. *IEEE/CVF International Conference on Computer Vision (ICCV), 2021*, 9660–9669.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. International conference on machine learning (pp. 1597–1607).

Chen, X., & He, K. (2021). Exploring simple siamese representation learning. *Proc. ieee conf. comput. vis. pattern recognit.* (pp. 15750–15758).

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629.

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 215–223).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (cvpr'05)* (vol. 1, pp. 886–893).

Datta, S., Sharma, G., & Jawahar, C. (2018). Unsupervised learning of face representations. In *2018 13th IEEE international conference on automatic face and gesture recognition (fg 2018)* (pp. 135–142).

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. *Proc. IEEE Conf. Comput. vis. pattern Recognit.* (pp. 4690–4699).

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., & Zisserman, A. (2021). With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *Proc. IEEE Conf. Comput. vis. Pattern Recognit.* (pp. 9588–9597).

Florea, C., Florea, L., Badea, M.-S., Vertan, C., & Racoviteanu, A. (2019). Annealed label transfer for face expression recognition. *Bmvc* (p. 104).

Gamble, J. A., & Huang, J. (2020). Convolutional neural network for human activity recognition and identification. In *2020 IEEE International Systems Conference (syscon)* (p. 1-7).

GE, C., Wang, J., Tong, Z., Chen, S., Song, Y., & Luo, P. (2023). Soft neighbors are positive supporters in contrastive visual representation learning. *The eleventh international conference on learning representations.*

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In Neural information processing: 20th international conference, daegu, korea. Proceedings, Part III 20 (pp. 117–124). Springer berlin heidelberg.

Harini, R., & Chandrasekar, C. (2012). Image segmentation using nearest neighbor classifiers based on kernel formation for medical images. *International conference on pattern recognition, informatics and medical engineering (prime-2012)* (p. 261-265).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proc. IEEE Conf. Comput. vis. Pattern Recognit.* (pp. 9729–9738).

Jakab, T., Gupta, A., Bilen, H., & Vedaldi, A. (2018). Unsupervised learning of object landmarks through conditional image generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems (vol. 31)*. Newry: Curran Associates, Inc.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Handbook Systemic Autoimmune Diseases, 1*(4), 66.

Li, W., Abtahi, F., Zhu, Z., & Yin, L. (2017). Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE international conference on automatic face and gesture recognition (fg 2017)* (pp. 103–110).

Li, Y., & Shan, S. (2023). Contrastive learning of person-independent representations for facial action unit detection. *IEEE Transactions on Image Processing, 32*, 3212–3225.

Li, Y., Zeng, J., & Shan, S. (2020). Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 302–317.

Li, Y., Zeng, J., & Shan, S. (2022). Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 302–317.

Li, Y., Zeng, J., Shan, S., & Chen, X. (2019). Self-supervised representation learning from videos for facial action unit detection. In *Proc. IEEE Conf. Comput. vis. Pattern Recognit.* (pp. 10924–10933).

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In *Proc. IEEE Conf. Comput. vis. Pattern Recognit.* (pp. 1871–1880).

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proc. IEEE Conf. Comput. vis. Pattern Recognit.* (pp. 212–220).

Liu, Y., Dai, W., Fang, F., Chen, Y., Huang, R., Wang, R., & Wan, B. (2021). Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition. *Information Sciences, 578*, 195–213.

Liu, Y., Wang, W., Zhan, Y., Feng, S., Liu, K., & Chen, Z. (2023, June). Pose-disentangled contrastive learning for self-supervised facial representation. In *Proc. IEEE Conf. Comput. vis. Pattern Recognit.* (pp. 9717-9728).

Lu, L., Tavabi, L., & Soleymani, M. (2020). Self-supervised learning for facial action unit recognition through temporal consistency. *British machine vision conference.*

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., & Bovik, A. C. (2022). Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing, 31*, 4149–4161.

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

McCann, S., & Lowe, D. G. (2012). Local naive bayes nearest neighbor for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3650–3656).

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A Large-Scale Speaker Identification Dataset. *Proc. interspeech 2017* (pp. 2616–2620). https://doi.org/10.21437/Interspeech.2017-950

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 971–987.

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. Bmvc 2015-proceedings of the british machine vision conference 2015.

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic cosine similarity. In *The 7th international student conference on advanced science and technology icast* (vol. 4, p. 1).

Roy, S., & Etemad, A. (2021). Self-supervised contrastive learning of multi-view facial expressions. In *ICMI - proc. int. conf. multimodal interact.* (pp. 253–257).

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 397–403).

Samanta, A., & Guha, T. (2017). On the role of head motion in affective expression. In *2017 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 2886–2890).

Shao, Z., Liu, Z., Cai, J., & Ma, L. (2018). Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 705–720).

Shu, Y., Gu, X., Yang, G.-Z., & Lo, B. P. L. (2022). Revisiting self-supervised contrastive learning for facial expression recognition. In *33rd British machine vision conference 2022, BMVC 2022, London, November 21–24, 2022.* BMVA Press.

Shu, Z., Sahasrabudhe, M., Guler, R. A., Samaras, D., Paragios, N., & Kokkinos, I. (2018). Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the european conference on computer vision (ECCV)* (pp. 650–665).

Wiles, O., Koepke, A. S., & Zisserman, A. (2018). Self-supervised learning of a facial attribute embedding from video. British machine vision conference.

Yang, S., Wang, Y., van de Weijer, J., Herranz, L., & Jui, S. (2021). Exploiting the intrinsic neighborhood structure for source-free domain adaptation. CoRR, arXiv:2110.04202.

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. (2006). A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (fgr06)* (pp. 211-216).

Yu, J., Zhou, H., Zhan, Y., & Tao, D. (2021). Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In*Proceedings of the aaai conference on artificial intelligence* (vol. 35, pp. 4626–4634).

Zhang, R., Isola, P., & Efros, A. A. (2017). Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 1058–1067).

Zhao, K., Chu, W.-S., & Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 3391–3399).

Zhao, S., Cai, H., Liu, H., Zhang, J., & Chen, S. (2018). Feature selection mechanism in cnns for facial expression recognition. *Bmvc* (vol. 12, p. 317).

Zhao, X., Shi, X., & Zhang, S. (2015). Facial expression recognition via deep learning. *IETE Technical Review, 32*(5), 347–355.

Zheng, T., & Deng, W. (2018). Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, *5*(7).

Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In *2016 IEEE conference on computer vision and pattern recognition (cvpr)* (pp. 146–155). https://doi.org/10.1109/CVPR.2016.23