Pose-disentangled Contrastive Learning for Self-supervised Facial Representation

Yuanyuan Liu[1*], Wenbin Wang[1*], Yibing Zhan[2], Shaoze Feng[1], Kejun Liu[1], Zhe Chen[3†]
[1]School of Computer Science, China University of Geosciences, Wuhan, China
[2]JD Explore Academy, China
[3]The University of Sydney, Australia
{liuyy, wangwenbin, fengshaoze, liukejun}@cug.edu.cn; zhanyibing@jd.com; zhe.chen1@sydney.edu.au

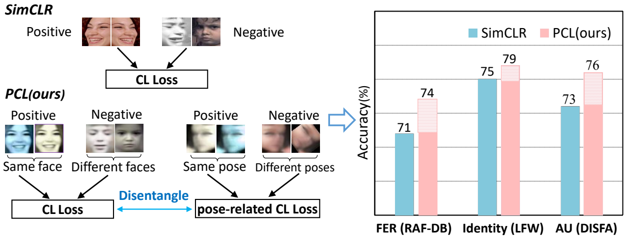
Abstract

Self-supervised facial representation has recently attracted increasing attention due to its ability to perform face understanding without relying on large-scale annotated datasets heavily. However, analytically, current contrastive-based self-supervised learning (SSL) still performs unsatisfactorily for learning facial representation. More specifically, existing contrastive learning (CL) tends to learn pose-invariant features that cannot depict the pose details of faces, compromising the learning performance. To conquer the above limitation of CL, we propose a novel Pose-disentangled Contrastive Learning (PCL) method for general self-supervised facial representation. Our PCL first devises a pose-disentangled decoder (PDD) with a delicately designed orthogonalizing regulation, which disentangles the pose-related features from the face-aware features; therefore, pose-related and other pose-unrelated facial information could be performed in individual subnetworks and do not affect each other's training. Furthermore, we introduce a pose-related contrastive learning scheme that learns pose-related information based on data augmentation of the same image, which would deliver more effective face-aware representation for various downstream tasks. We conducted linear evaluation on four challenging downstream facial understanding tasks, i.e., facial expression recognition, face recognition, AU detection and head pose estimation. Experimental results demonstrate that PCL significantly outperforms cutting-edge SSL methods. Our Code is available at https://github.com/DreamMr/PCL.

1. Introduction

Human face perception and understanding is an important and long-lasting topic in computer vision. By analyzing faces, we can obtain various kinds of information, including identities, emotions, and gestures. Recently, deep convolutional neural networks (DCNNs) [20, 30, 62] have achieved promising facial understanding results, but they require a large amount of annotated data for model training. Since labeling face data is generally a labor- and time-costly process [61], it becomes important to enable DCNN models to learn from unlabelled face images, which are much easier to collect. Accordingly, researchers have introduced self-supervised learning (SSL) schemes to achieve better learning performance on unlabeled facial data.

To achieve effective SSL performance, contrastive learning (CL) based strategy is widely applied in the community [6,26,43]. In general, a CL-based method pulls two features representing similar samples closer to each other and pushes those of diverse samples far away from each other [56], thus facilitating the DCNNs to learn various visual patterns without annotations. Generally, without supervision, similar/positive samples of CL are obtained by augmenting the same image, and the diverse/negative samples can refer to different images. To learn from unlabelled face images, ex-



(a) Motivation of SimCLR and our PCL (b) Results of SimCLR and our PCL

Figure 1. The motivation of our method. Affected by different poses, the popular CL method, *e.g.*, SimCLR, treats pose and other face information uniformly, resulting in sub-optimal results. To alleviate this limitation for CL, our PCL attempts to disentangle the learning on pose-related features and pose-unrelated facial features, thus achieving more effective self-supervised facial representation learning for downstream facial tasks.

[*] Equally-contributed first authors
[†] Corresponding author

isting CL-based methods [48,53,65] have achieved effective self-supervised facial representation learning.

However, despite progress, we found that directly utilizing CL-based methods still obtained sub-optimal performance due to the facial poses. In particular, CL-based methods treat the augmented images from the same image as positive samples. In such a manner, the learned features are pose-invariant, which cannot recognize the variances of facial poses. Nevertheless, poses are one significant consideration for facial understanding [1, 51]; for example, a person tends low their head when they feel sad.

To tackle the above limitation of CL, we propose a Pose-disentangled Contrastive Learning (PCL) method, which disentangles the learning on pose-related features and pose-unrelated facial features for CL-based self-supervised facial representation learning. Fig. 1 has shown an intuitive example of contrastive learning results. Specifically, Our method introduces two novel modules, *i.e.*, a pose-disentangled decoder (PDD) and a pose-related contrastive learning scheme (see Fig. 2). In the PDD, we first obtain the face-aware features from a backbone, such as ResNet [10, 27], Transformer [11, 16–18, 40], and then disentangle pose-related features and pose-unrelated facial features from the face-aware features using two different subnets through facial reconstruction. In facial reconstruction, the combination of one pose-unrelated facial feature and one pose-related feature can reconstruct an image with the same content as the pose-unrelated facial feature and the same pose as the pose-related feature. Furthermore, an orthogonalizing regulation is designed to make the pose-related and pose-unrelated features more independent.

In the pose-related contrastive learning, instead of learning pose-invariant features by normal CL, we introduce two types of data augmentation for one face image, one containing pose augmentation and another only containing pose-unrelated augmentation. Therefore, image pairs generated by using pose augmentation contain different poses and serve as negative pairs, whereas image pairs generated from pose-unrelated augmentation contain the same pose as the original image and are treated as positive pairs. The pose-related CL is conducted to learn pose-related features, and face CL is used to learn pose-unrelated facial features. Therefore, our proposed pose-related CL can learn detailed pose information without disturbing the learning of pose-unrelated facial features in the images.

In general, the major contributions of this paper are summarized as follows:

1. We propose a novel pose-disentangled contrastive learning framework, termed PCL, for learning unlabeled facial data. Our method introduces an effective mechanism that could disentangle pose features from facial features and enhance contrastive learning for pose-related facial representation learning.
2. We introduce a PDD using facial image reconstruction with a delicately designed orthogonalizing regulation to help effectively identify and separate the face-aware features obtained from the backbone into pose-related and pose-unrelated facial features. The PDD is easy-to-implement and efficient for head pose extraction.
3. We further propose a pose-related contrastive learning scheme for pose-related feature learning. Together with face contrastive learning on pose-unrelated facial features, we make both learning schemes cooperate with each other adaptively and obtain more effective learning performance on the face-aware features.
4. Our PCL can be well generalized to several downstream tasks, *e.g.*, facial expression recognition (FER), facial AU detection, facial recognition, and head pose estimation. Extensive experiments show the superiority of our PCL over existing SSL methods, accessing state-of-the-art performance on self-supervised facial representation learning.

2. Related Work

Contrastive Learning Contrastive learning (CL) has been widely used in self-supervised learning and has yielded significant results in many vision tasks [6–9, 25, 26, 46, 57, 66]. CL aims to map features of samples onto a unit hypersphere such that the feature distances of the positive sample pairs on the sphere are similar. In contrast, the feature distances of the randomly sampled negative sample pairs are pushed apart [56]. Recent breakthroughs in CL, such as MoCo [26] and SimCLR [6], shed light on the potential of discriminative models for visual representation. Thanks to a large number of negative samples, MoCo maintained a queue of negative samples to improve the capacity of CL [26]. Chen *et al.* [6] proposed a new self-supervised framework SimCLR to model the similarity of two images for learning visual representations without human supervision. SimSiam is proposed for exploring simple siamese representation learning by maximizing the similarity between two augmentations of one image, subject to certain conditions for avoiding collapsing solutions [8].

Self-supervised Facial Representation Learning Self-supervised facial representation learning is important for many face-related applications, such as FER, face recognition, AU detection, etc. [3–5, 28, 31, 35, 58]. Due to its capacity of learning on unlabelled data, an increasing number of research efforts are focusing on self-supervised face representation learning. FAb-Net used the motion changes between different frames of a video to learn facial motion features, and has achieved good results in FER [31]. Li *et al.* [34, 35] proposed a Twin-Cycle Autoencoder that

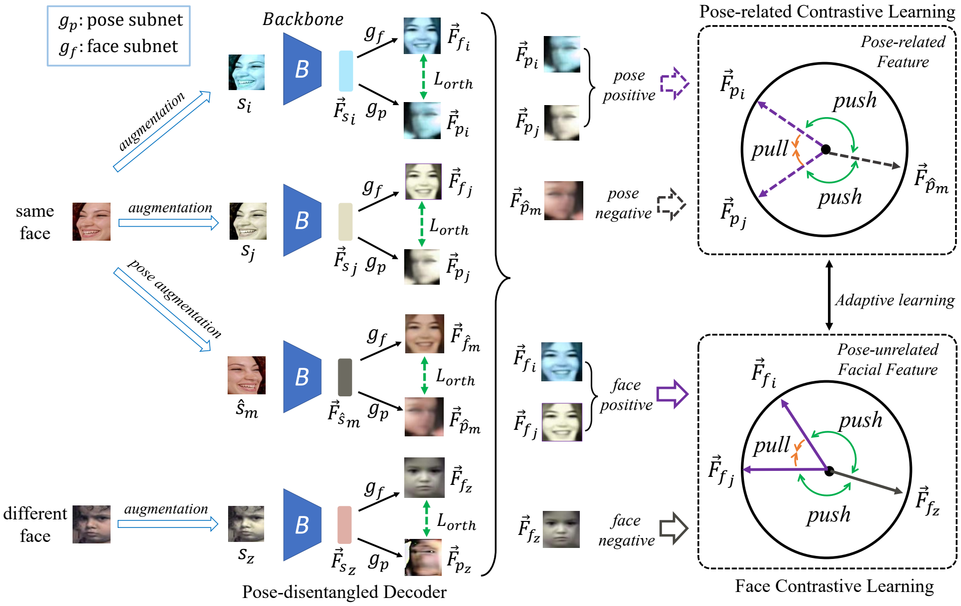


Figure 2. The overview of PCL for self-supervised facial representation learning. We first use a pose-disentangled decoder with an orthogonalizing regulation L_{orth} to help extract pose-related features (*e.g.*, $\vec{F}_{pi}$) and pose-unrelated facial features (*e.g.*, $\vec{F}_{fi}$) from input augmented images (*e.g.*, s_i), and then introduce pose-related contrastive learning and face contrastive learning schemes to further learn on the extracted features adaptively, resulting in more effective face-aware representation learning.

can disentangle the facial action-related movements and the head motion-related ones, obtaining good facial emotion representation for self-supervised AU detection [35]. FaceCycle decoupled facial expression and identity information via cyclic consistency learning to extract robust unsupervised facial representation, thus achieving good results in both FER and facial recognition [3]. Zheng *et al.* presented an study about the transferable visual models learned in a visual-linguistic manner on general facial representations [65]. Roy *et al.* [48] proposed a CL-MEx for pose-invariant expression representation by exploiting facial images captured from different angles. Shu *et al.* [53] used three sample mining strategies in CL to learn expression-related features. Overall, most of the existing work is crafting facial representation learning for a single task, and general self-supervised facial representation learning remains an open research problem.

3. The Proposed Approach

The overview of our proposed pose-disentangled contrastive learning is presented in Fig. 2. Our PCL mainly consists of two novel modules, *i.e.*, a pose-disentangled decoder (PDD) and a pose-related contrastive learning scheme. Tacking a face image as input, the PDD of PCL first employs a backbone network like ResNet [10,27] to extract general facial features and then attaches two subnets to produce separate pose-related features and pose-unrelated facial features. To train the PDD properly, we reconstruct the face through the combination of the two types of features with an orthogonalizing regulation posed on the separated features for better disentanglement. Then, we introduce pose-related contrastive learning to train the pose-related features and use face contrastive learning scheme to learn pose-unrelated facial features. We make the two learning objectives cooperate with each other adaptively, obtaining more promising self-supervised facial representations. Our PCL method can fulfill the training of neural networks in an end-to-end manner. In the following sections, we will describe the details of PCL.

3.1. Pose-disentangled Decoder

Previous CL-based methods [6, 26] treat pose and other facial information uniformly, resulting in pose-invariant features that cannot recognize the details of poses. One possible solution is not using pose augmentation for training. However, such a manner would reduce the training data diversity and further reduce the performance. To conquer the above limitation of CL, we design a PDD to disentangle the pose-related and pose-unrelated facial representations from the face-aware features. Therefore, through the individual learning information from the pose-related and pose-unrelated facial features, the face-aware features could be used as a well facial representation that properly depicts the face, including the pose and other useful information.

Nevertheless, identifying and separating the pose-related and pose-unrelated facial features is nontrivial. To tackle this problem, the PDD assists the training of the two types

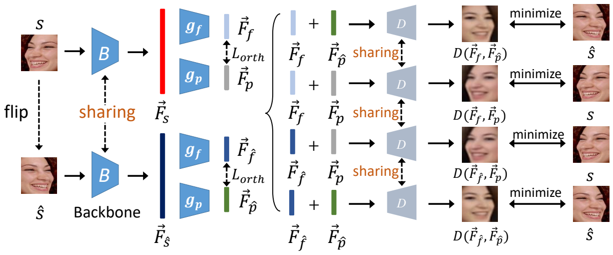

s
$\hat{s}$
flip
B
sharing
B
Backbone
$\vec{F}_s$
$\vec{F}_{\hat{s}}$
g_f
g_p
g_f
g_p
$\vec{F}_f$
L_{orth}
$\vec{F}_p$
$\vec{F}_{\hat{f}}$
L_{orth}
$\vec{F}_{\hat{p}}$
$\vec{F}_f$ + $\vec{F}_{\hat{p}}$
$\vec{F}_f$ + $\vec{F}_p$
$\vec{F}_{\hat{f}}$ + $\vec{F}_p$
$\vec{F}_{\hat{f}}$ + $\vec{F}_{\hat{p}}$
sharing
sharing
sharing
D
D
D
D
$D(\vec{F}_f, \vec{F}_{\hat{p}})$
$D(\vec{F}_f, \vec{F}_p)$
$D(\vec{F}_{\hat{f}}, \vec{F}_p)$
$D(\vec{F}_{\hat{f}}, \vec{F}_{\hat{p}})$
minimize
minimize
minimize
minimize
$\hat{s}$
s
s
$\hat{s}$

Figure 3. The training pipeline of PDD. Given the face image s and its pose-varied image $\hat{s}$ as input, we first use a backbone to encode facial features of input images, and then use two separating subnets, *i.e.*, $g_f(\cdot)$ and $g_p(\cdot)$, to extract pose-unrelated and pose-related facial features, respectively. Finally, we employ a face reconstruction network D to translate the extracted two types of features into reconstructed faces. Moreover, an orthogonalizing regulation L_{orth} is used for training the PDD to make the separated features independent of each other.

of features through reconstructing faces: one image with a specific pose could be reconstructed through the combination of the pose-unrelated facial feature of the corresponding image and the pose-related feature of the given pose.

The overall pipeline of PDD is presented in Fig. 3. PDD consists of a shared backbone network, two separating subnet branches, and a shared reconstruction network. In this paper, the backbone network is a shallow 16-layer residual network for learning facial features from input face image. Note that our PCL can marry with any other backbones, such as VGG [55] and Transformer. Then, two extra separating subnets attach to the backbone are used to separate the obtained features into the pose-related features and the pose-unrelated facial features (both features are 2048-dimensional features in practice), respectively. Finally, we employ a 6-layer blocks with each of an upsampling layer and convolutional layer as the reconstruction network to translate the combination of the pose-related features and pose-unrelated facial features into a reconstructed face.

Formally, given an input face image s, we represent its pose as p. We represent the same face with a different pose as $\hat{s}$ with its pose $\hat{p}$. In PDD, we use the backbone B to encode face data into a facial feature $\vec{F}_s$ with the pose. We would like to mention again that the final learned $\vec{F}_s$ is the self-supervised face-aware representation for the downstream task evaluation. Then, two separating branches, denoted as $g_p(\cdot)$ and $g_f(\cdot)$, are employed to extract the pose-related feature $\vec{F}_p$ and the pose-unrelated facial feature $\vec{F}_f$, respectively. Meanwhile, using the same backbone and separating branches, we have corresponding features $\vec{F}_{\hat{f}}$ and $\vec{F}_{\hat{p}}$ for the pose-varied face $\hat{s}$. According to the goal of PDD, the $\vec{F}_f$ and $\vec{F}_{\hat{f}}$ are supposed to represent the same facial features, while the $\vec{F}_p$ and $\vec{F}_{\hat{p}}$ are supposed to describe different pose-related features. To achieve this, we introduce a reconstruction network D to translate pose-related and pose-unrelated facial features into reconstructed faces that can be defined explicitly. As a result, we suppose the PDD should satisfy the following transformations:

$$\begin{aligned} D(\vec{F}_f, \vec{F}_p) = s,\ D(\vec{F}_f, \vec{F}_{\hat{p}}) = \hat{s}, \\ D(\vec{F}_{\hat{f}}, \vec{F}_p) = s,\ D(\vec{F}_{\hat{f}}, \vec{F}_{\hat{p}}) = \hat{s}. \end{aligned} \tag{1}$$

The above goals indicate that the PDD should reconstruct the same face but different poses according to varied pose-related features. When the above transformations can be satisfied, we can then consider that the PDD tends to have the ability to separate the pose-related feature F_p from pose-unrelated facial feature F_f properly. Otherwise, for example, if $\vec{F}_f$ still contains redundant feature about p, the $D(\vec{F}_f, \vec{F}_{\hat{p}})$ would not generate the $\hat{s}$ image appropriately and would tend to produce the s instead.

To make PDD satisfy the above transformations, the disentangled objective L_{dis} of the PDD is:

$$\begin{aligned} L_{dis} = &||s - D(\vec{F}_f, \vec{F}_p)||_1 + ||\hat{s} - D(\vec{F}_f, \vec{F}_{\hat{p}})||_1 \\ &+ ||s - D(\vec{F}_{\hat{f}}, \vec{F}_p)||_1 + ||\hat{s} - D(\vec{F}_{\hat{f}}, \vec{F}_{\hat{p}})||_1, \end{aligned} \tag{2}$$

where $|| \cdot ||_1$ represents l_1-norm. Additionally, we also try to use GAN [21] instead of the l_1-norm; however, GAN can only make generated images approximate the real images but cannot guarantee the poses of the generated images. For more discussion, see our supplemental material.

Moreover, to disentangle the extracted features more properly, an orthogonalizing regulation is further introduced to make the extracted features uncorrelated. Therefore, we constrain that the $\vec{F}_f$ and $\vec{F}_p$ should be orthogonal to each other. To achieve this, inspired by [2,24,36,49], the orthogonalizing regulation L_{orth} is defined as follows:

$$L_{orth} = \frac{1}{N}\left(\sum_{i=1}^{N} ||\vec{F}_f \cdot \vec{F}_p||_2^2 + \sum_{i=1}^{N} ||\vec{F}_{\hat{f}} \cdot \vec{F}_{\hat{p}}||_2^2\right). \tag{3}$$

During learning, minimizing the L_{orth} can help force the dot-products of pose-related and pose-unrelated facial features to reach near zero, thus making them orthogonal to each other. Finally, we define the total optimization objective L_{PDD} of the PDD, including the disentangled loss and orthogonalizing regulation as:

$$L_{PDD} = L_{orth} + L_{dis}. \tag{4}$$

3.2. Pose-related Contrastive Learning

Normal CL tends to learn pose-invariant features. Therefore, we further devise a Pose-related Contrastive Learning to enable effective self-supervised learning on pose information, suppressing the side effects of pose-invariant features. Since it is unknown whether different faces have the same pose or not, it is difficult to construct pose positive and negative sample pairs well by directly using data augmentation in contrastive learning. To address this problem, unlike the normal CL that treats different face individuals as negative pairs, we propose a pose augmentation method for pose-related contrastive learning, *i.e.*, for the same face image, we apply pose transformation and image transformation to it separately, then consider the same image pairs containing different (augmented) poses as negative pairs and the same image with containing the same (unaugmented) pose as positive pairs for contrastive learning on the pose. Through this way, pose-related contrastive learning can focus on learning pose information without being influenced by images with the same pose as the negative samples.

Formally, for the input face image s, we use the specific pose augmentation (such as flipping and rotation) to generate M negative samples $\hat{s}_M$, *i.e.*, $\hat{s}_M = \{\hat{s}_m\}_{m=1}^{M}$, resulting in a negative pair ($\hat{s}_i$ and $\hat{s}_m$), while using a stochastic data augmentation (such as, random crop, color jitter, Gaussian blur, and Sobel filtering) to obtain a positive pair (s_i and s_j), as shown in Fig. 2. Both the positive and negative pairs are passed through the PDD to extract the pose-related features as $\vec{F}_{p_i}, \vec{F}_{p_j}$ and $\vec{F}_{\hat{p}_m}$, respectively. Overall, the pose-related contrastive loss is written as:

$$L_{pose}(\vec{F}_{p_i}, \vec{F}_{p_j}, \vec{F}_{\hat{p}_m}) = l_p(\vec{F}_{p_i}, \vec{F}_{p_j}) + l_p(\vec{F}_{p_j}, \vec{F}_{p_i}),$$

$$l_p(\vec{F}_{p_i}, \vec{F}_{p_j}) = -log\frac{exp(\frac{sim(\vec{F}_{p_i}, \vec{F}_{p_j})}{\tau})}{\sum_{m=1}^{M} exp(\frac{sim(\vec{F}_{p_i}, \vec{F}_{\hat{p}_m})}{\tau})}, \tag{5}$$

where $sim(\cdot)$ is the pairwise cosine similarity. τ denotes a temperature parameter. Through the pose-related contrastive learning, our PCL can learn more detailed pose information from facial images without disturbing the learning of pose-unrelated facial features.

3.3. Overall Optimization Objectives

Together with pose-related contrastive learning on the pose-related features, we employ face contrastive learning on the pose-unrelated facial features, thus using two different subnetworks with different CL strategies to alleviate the side effects of pose information for learning face patterns. More specifically, we randomly sample a minibatch of N face images, and use a stochastic data augmentation to transform any given input face image s, resulting in two correlated views of the same face as a positive pair s_i and s_j. Secondly, each positive pair, *e.g.*, s_i and s_j in Fig. 2, are passed through the PDD to extract the pose-unrelated facial features $\vec{F}_{f_i}$ and $\vec{F}_{f_j}$, respectively. The contrastive loss on the face branch is written as:

$$L_{face}(\vec{F}_{f_i}, \vec{F}_{f_j}) = l_f(\vec{F}_{f_i}, \vec{F}_{f_j}) + l_f(\vec{F}_{f_j}, \vec{F}_{f_i}),$$

$$l_f(\vec{F}_{f_i}, \vec{F}_{f_j}) = -log\frac{exp(\frac{sim(\vec{F}_{f_i}, \vec{F}_{f_j})}{\tau})}{\sum_{z=1}^{2N} 1_{[i\neq z]} exp(\frac{sim(\vec{F}_{f_i}, \vec{F}_{f_z})}{\tau})}, \tag{6}$$

where $\vec{F}_{f_z}$ is from negative pairs.

Therefore, during training, our PCL has three major objectives: the disentangled lose L_{PDD} of PDD, the pose-related contrastive loss L_{pose} on the pose-related features, and the face contrastive loss L_{face} on the pose-unrelated facial features. The overall objective function L of the PCL is the weighted sum of L_{PDD}, L_{pose}, and L_{face}. Mathematically, the total loss L can be written as:

$$L = L_{PDD} + \alpha_{pose} \cdot L_{pose} + \alpha_{face} \cdot L_{face}, \tag{7}$$

where α_{pose} and α_{face} are two dynamic weights to adaptively balance the pose and face learning objectives in the multi-task learning manner according to their contributions to facial representations. We employ the Dynamic Weight Average (DWA) [37] to obtain the α_{pose} and α_{face} during training. More details of dynamic weight learning can be seen in the supplemental material. We also show in the experiments that adding the dynamic weight learning improves performance (see Table 5), demonstrating the usefulness of adaptive cooperation of two CL schemes.

4. Experiments

In this section, we verified the effectiveness of our proposed PCL by answering two questions:

Q1: does our facial representation perform well and has generalizability? (Refer to section 4.2)

Q2: whether the improvements come from the contributions we proposed in this paper? (Refer to section 4.3)

We further visualized the contents of learned features to demonstrate the reasonability of PCL. (Refer to section 4.4)

4.1. Experimental Settings

Datasets The proposed PCL was trained on the combination of VoxCeleb1 [44] and VoxCeleb2 [12] datasets without any annotations. The VoxCeleb1 and VoxCeleb2 have 299,085 video clips of around 7,000 speakers. We extracted the video frames at 6 fps, cropped to faces shown in the center of frames and then resized to the resolution of 64 × 64 for training [3].

For FER evaluation, we used two widely-used FER datasets, *i.e.*, FER-2013 [22] and RAF-DB [32]. The FER-2013 consists of 28,709 training and 3,589 testing images. We followed the experimental setup as [3] to particularly use the basic emotion subset of RAF-DB with 12,271 training and 3,068 testing images.

For facial recognition evaluation, we adopt two in-the-wild facial identity datasets, *i.e.*, LFW [29] and CPLFW [64]. The LFW consists of 13,233 face images from 5,749 identities and has 6,000 face pairs for evaluating identity verification. The CPLFW dataset includes 3,000 positive face pairs with pose differences to add pose variation to intra-class variance. All reported results were averaged across the 10 folds.

For facial AU detection, we evaluated our method on the DISFA [42] dataset with 26 participants. The AUs are labeled with intensities from 0 to 5. The frames with intensities greater than 1 are considered positive, while others are treated as negative. In total, we obtained about 130,000 AU-labelled frames and followed the experimental setup of [35] to conduct a 3-fold cross-validation.

For head pose estimation, we adopt two widely-used tasks, *i.e.*, pose regression (trained on 300W-LP [50] and evaluated on AFLW2000 [67]) and pose classification (on BU-3DFE [59]). The 300W-LP contains 122,450 images and AFLW2000 contains 2000 images. For pose classification, following the experimental setup as [39], we used BU-3DFE with 14,112 images for training and the rest of 6,264 images for validation.

Implementation Details Our proposed model was implemented based on the PyTorch framework and trained with the Adam optimizer ($\beta_1 = 0.9$, and $\beta_2 = 0.999$) for 1000 epochs. The batch size and initial learning rate are set to 256 and 0.0001, respectively. The learning rate is decreased by cosine annealing. The temperature parameter τ is set to 0.07. The baseline SimCLR [6] used the data augmentation (such as random crop, color jitter, Gaussian blur, and Sobel filtering) and negative interpolation [66] for training.

Referring to [3] and [31], the backbone of our model is a simple 16-layer CNN, and the reconstruction network is a simple 6-layer block with each of an upsampling layer and a convolutional layer. The $g_f(\cdot)$ and $g_p(\cdot)$ are convolutional subnets with the same architecture. We will give the detailed network structure in the supplementary material.

In addition, we explored different choices of varying the pose p for training PDD like flipping and rotation. However, the experimental results demonstrate that flipping p is the most effective way to help PDD learn to identify and separate pose from facial representations (0.43% improvement over adding rotation and translation). To trade off between efficiency and accuracy, we used pose flipping for training PDD in this study.

Evaluation Protocols We followed the widely used linear evaluation protocol in SSL [3, 6, 8, 9, 14, 23, 25, 26, 35] to verify our method. The linear classifier is a simple linear fully-connected layer, and is trained with the frozen self-supervised face-aware representation $\vec{F}_s$ from the backbone B for 300 epochs.

Following [3, 14, 35], we resized the images to the size 100×100, 128×128, 256×256 and 256×256 respectively, for FER, face recognition, AU detection and pose-related downstream tasks.

4.2. Performance Comparison for Q1

4.2.1 Evaluation for Facial Expression Recognition

Given the trained model, we investigated the learned $\vec{F}_s$ by evaluating the performance of its applications on FER. The quantitative results shown in Table 1 demonstrate that our proposed method is able to provide superior performance with respect to other methods. Compared to the SimCLR [6], the proposed PCL improves the accuracy by over 7.3% and 3.41%, respectively. These results suggest that our PCL can be used as a pretext task to learn an effective self-supervised facial representation with rich expression information for the FER task.

Table 1. Evaluation of the FER task on the FER-2013 and RAF-DB datasets. (Note: the highest results of self-supervised methods are highlighted in bold, and * indicates the results reproduced by authors.)

	FER-2013	RAF-DB
Method	Accuracy(%)	Accuracy(%)
Fully supervised		
FSN [63]	67.60	81.10
ALT [19]	69.85	84.50
Self-supervised (linear evaluation)		
LBP [45]	37.89	52.17
HoG [13]	45.47	63.53
FAb-Net [31]	46.98	66.72
TCAE [35]	45.05	65.32
BMVC'20 [41]	47.61	58.86
MoCo [26]	47.24	68.32
FaceCycle [3]	48.76	71.01
SimCLR [6]*	49.51	71.06
Ours	**56.81**	**74.47**

4.2.2 Evaluation for Facial Recognition

For the facial recognition task, our learned self-supervised face-aware features also outperform other self-supervised-based facial representations. As shown in Table 2, our PCL achieved the best accuracy of 79.72% and 64.61% on LFW and CPLFW, respectively, which are 3.75% and 1.26% better than the results of the state-of-the-art method. The improvements suggest that our PCL can be used as an effective pretext task for facial identity recognition.

4.2.3 Evaluation for Facial AU Detection

Facial AU detection estimates whether each AU in the face image or video is activated. We followed the [35] and used a binary cross-entropy loss to train a linear classifier for AU detection. Table 3 reports the comparison of our PCL and the state-of-the-art self-supervised methods, as well as the full supervised methods. We evaluated not only the same backbone approaches as ours but also deeper backbone approaches. The results show that our method still has a clear advantage. As shown in Table 3, our method outperforms other self-supervised methods in the average F1 score. Thanks to disentangled facial features, the learned facial representation can better reflect facial actions. In addi-

Table 2. Evaluation of facial recognition on the LFW and CPLFW datasets. (Note: the highest results of self-supervised methods are highlighted in bold, and * indicates the results reproduced by authors.)

	LFW	CPLFW
Method	Accuracy(%)	Accuracy(%)
Fully supervised		
VGG-Face [47]	98.95	84.00
SphereFace [38]	99.42	81.40
ArcFace [15]	99.53	92.08
Self-supervised (Linear evaluation)		
LBP [45]	72.44	-
VGG [14]	72.20	-
MoCo [26]*	65.88	57.82
SimCLR [6]*	75.97	62.25
FaceCycle [3]*	74.12	63.35
Ours	**79.72**	**64.61**

tion, our PCL has reached the fully supervised level, and the average $F1$ has exceeded the full supervised DRML [62] by 28.1 and the EAC-Net [33] by 6.3, respectively.

Table 3. Evaluation of facial AU detection on the DISFA dataset. We use $F1$ score for the evaluation. (Note: the highest results of self-supervised methods are highlighted in bold, and * indicates the results reproduced by authors.)

	Methods/AU	1	2	4	6	9	12	25	26	ave
Supervised	DRML [62]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
	EAC-Net [33]	41.5	26.4	66.4	50.7	80.5	89.3	88.9	15.6	48.5
	JAA-Net [52]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
Self-superivised	SplitBrain [60]	13.1	10.6	35.7	40.2	30.2	57.5	77.4	40.3	38.1
	DeformAE [54]	17.6	12.3	46.7	43.5	26.0	62.7	64.8	**47.6**	40.1
	Fab-Net [31]	15.5	16.2	43.2	**50.4**	23.2	69.6	72.4	42.4	41.6
	TCAE [35]	15.1	16.2	50.5	48.7	23.3	72.1	72.4	42.4	45.0
	TCAE [35]*	10.5	13.3	20.9	18.8	7.5	44.7	57.8	9.9	22.9
	FaceCycle [3]*	26.4	10.2	37.3	21.5	25.0	71.8	84.2	34.7	38.9
	SimCLR [6]*	40.5	46.9	53.8	33.5	24.9	74.7	85.0	35.6	49.4
	Ours	**53.8**	**44.9**	**58.1**	37.2	**53.2**	**73.1**	**86.5**	31.3	**54.8**

4.2.4 Evaluation for Head Pose Estimation

We evaluated our PCL on two pose-related tasks, including pose regression (trained on 300W-LP and evaluated on AFLW2000) and pose classification (on BU-3DFE). We compared with different SSL methods in Table. 4. Our PCL achieved the lowest mean absolute error (MAE) of 12.36 on AFLW2000 and the best accuracy of 98.95% on BU-3DFE, outperforming the other self-supervised methods.

Table 4. Evaluation on head pose estimation. (↓ represents the smaller is better. ↑ represents the larger is better.)

	AFLW2000 (pretrained on 300W-LP)				BU-3DFE
	Yaw↓	Pitch↓	Roll↓	MAE↓	Accuracy (%)↑
FaceCycle [3]	11.70	**12.76**	12.94	12.47	98.82
MoCo [26]	28.49	16.29	15.55	20.11	75.33
SimCLR [6]	11.20	19.86	12.08	14.38	98.85
Ours	**9.86**	16.59	**10.62**	**12.36**	**98.95**

Table 5. Ablation study of the proposed PCL. Impact of integrating different components (*i.e.*, PDD and pose-related contrastive learning L_{pose}) into the baseline (SimCLR) on the RAF-DB dataset.

Baseline (SimCLR)	PDD		Contrastive learning		FER
	L_{dis}	L_{orth}	L_{pose}	Dynamic weighting	
✓					71.06
✓	✓				71.47
✓	✓	✓			72.39
✓	✓	✓	✓		73.73
✓	✓	✓	✓	✓	**74.47**

4.3. Ablation Study and Analysis for Q2

Effect of Different Modules To better understand the role of each module in our PCL, Table 5 presents the ablation results of the gradual addition of different components into the baseline (SimCLR w/o pose augmentation) for FER on the RAF-DB dataset. The baseline achieved a FER accuracy of 71.06%. Compared with the baseline, separating the pose-related features from face-aware features slightly improved the performance by 0.41%. The further addition of L_{orth} improved the FER accuracy to 72.39%. We emphasized that this is the result of using two normal contrastiveing learning schemes on the two features separated by PDD. A significant improvement of 1.34% was obtained after adding the pose-related contrast learning L_{pose}, verifying that pose-related face information can help improve CL-based self-supervised facial representation performance. Additionally, the dynamic weighting achieved the best accuracy of 74.47%.

Table 6. The effects of poses on SimCLR and our PCL (w/o Dynamic weighting).

Tasks	SimCLR w/o pose	SimCLR w/ pose	PCL w/o pose	PCL w/ pose
FER(RAF-DB)	71.06	73.17	73.24	**73.73**
Pose estimation(BU-3DFE)	98.93	98.85	98.40	**98.95**

Effect of Poses on Contrastive Learning In order to further discuss the pose-invariant face features learned by SimCLR [6] and the pose-related face-aware features learned by our PCL, Table 6 shows the comparison of SimCLR with and without pose augmentation, as well as our PCL with and without pose-related contrastive loss L_{pose}, respectively, on the RAF-DB dataset. SimCLR w/ pose achieved a relative accuracy increase of 2.97% to SimCLR w/o pose on FER, while a relative decrease in pose estimation (about 0.09%). The result demonstrates that learning pose-invariant features can help improve CL performance.

In addition, PCL w/ pose achieved satisfied improvement in both FER (relative increase of 3.76%) and pose estimation (an increase of 0.02%), due to effectively exploring pose-unrelated facial and pose-related features. However, PCL w/o pose can not learn pose-related information, resulting in a slight decrease in both FER (decrease 0.49%)

Table 7. Linear evaluation with different face features. $\vec{F}_f + \vec{F}_p$ means to add the pose-related feature $\vec{F}_p$ with the pose-unrelated facial feature $\vec{F}_f$, and $\vec{F}_s$ represents the face-aware feature.

Different features	RAF-DB	LFW	DISFA
$\vec{F}_f$	73.04	78.55	54.30
$\vec{F}_p$	65.71	62.55	34.17
$\vec{F}_s$	**74.47**	**79.72**	54.78
$\vec{F}_f + \vec{F}_p$	73.53	79.10	**56.26**

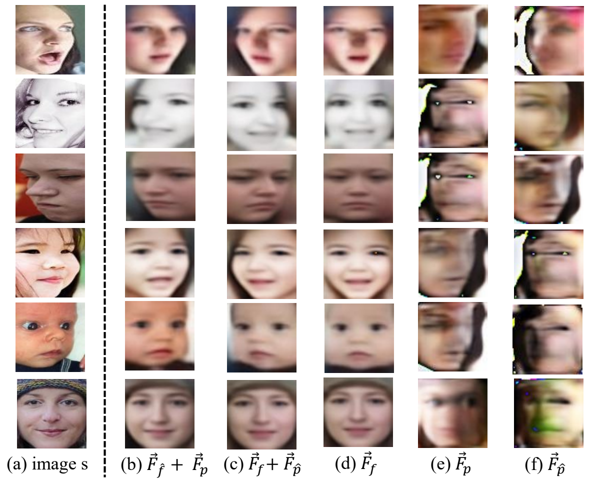

(a) image s (b) $\vec{F}_{\hat{f}} + \vec{F}_p$ (c) $\vec{F}_f + \vec{F}_{\hat{p}}$ (d) $\vec{F}_f$ (e) $\vec{F}_p$ (f) $\vec{F}_{\hat{p}}$

Figure 4. The reconstructed faces with disentangled pose-unrelated facial and pose-related features. (a) Source image s, (b)-(f) the reconstructed faces with different features. $\vec{F}_f$: pose-unrelated facial feature from s, $\vec{F}_p$: pose-related feature from s, $\vec{F}_{\hat{f}}$: pose-unrelated facial feature from pose-flipped $\hat{s}$, $\vec{F}_{\hat{p}}$: pose-related feature from pose-flipped $\hat{s}$.

and pose estimation (decrease 0.55%). The experiment result shows that poses are one significant consideration for facial understanding.

Comparison of Different Learned Features Table 7 shows a linear evaluation with different facial features extracted from the backbone and the followed two subnets, *i.e.*, $\vec{F}_s$ extracted from the backbone B, $\vec{F}_f$ extracted from $g_f(\cdot)$, and $\vec{F}_p$ extracted from $g_p(\cdot)$, in our PCL. For a fair comparison, the facial images were rescaled to the same size, and all the features were normalized to the same dimension in each case. The face-aware features $\vec{F}_s$ extracted from the backbone B achieved the best performance for FER and facial recognition tasks, respectively. Compared with single $\vec{F}_f$, We added the pose-related features $\vec{F}_p$ with the pose-unrelated facial features $\vec{F}_f$ and gained improvement on three tasks by 0.49%, 0.55%, and 1.96%, respectively. The result demonstrates that pose-related information can be complementary to face information for achieving more effective face-aware representation in CL.

4.4. Visualization

Fig. 4 visualizes the reconstructed faces with the pose-related and pose-unrelated facial features disentangled by our method. As shown in Fig. 4(b) and (c), our PCL successfully reconstructed the same faces but different poses according to varied pose-related features and the same pose-unrelated facial features, *i.e.* $\vec{F}_{\hat{f}} + \vec{F}_p$ and $\vec{F}_f + \vec{F}_{\hat{p}}$, which shows the capability in separating pose-related features. Fig. 4(d) shows the reconstructed frontal faces with the pose-unrelated facial features $\vec{F}_f$ from the image s, which demonstrates that our PCL is able to effectively disentangle the facial features without poses. Additionally, as shown in Fig. 4(e) and (f), we just used pose-related features from the image s and its pose-flipped image $\hat{s}$, *i.e.* $\vec{F}_p$ and $\vec{F}_{\hat{p}}$. Obviously, the generated images only include varied pose information with few face patterns.

5. Conclusions

In this paper, a novel pose-disentangled contrastive learning (PCL) is proposed for general self-supervised facial representation learning. PCL introduces two novel modules, *i.e.*, a pose-disentangled decoder (PDD) and a pose-related contrastive learning scheme. First, the PDD with a designed orthogonalizing regulation learns to disentangle pose-related features from face-aware features, thus obtaining pose-related and other pose-unrelated facial features independent of each other. Then, together with face contrastive learning on pose-unrelated facial features, we further propose a pose-related contrastive learning scheme on pose-related features. Both two learning schemes cooperate with each other adaptively for more effective self-supervised facial representation learning performance. With the two components, the proposed PCL achieved a vastly improved performance on four downstream face tasks, (*i.e.*, facial expression recognition, facial recognition, facial AU detection, and head pose estimation). Extensive experiments demonstrate that PCL is superior to other state-of-the-art self-supervised methods, obtaining strong robust self-supervised facial representation. In the future, we will continue to discuss the effects of other face-related attributions, such as ages, makeup and occlusion. We believe the proposed approach can be well extended to decouple other relevant information for more robust self-supervised and unsupervised facial representation.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China grant (Grant No. 62076227, 62002090) and Wuhan Applied Fundamental Frontier Project under Grant (No. 2020010601012166). Dr Zhe Chen is supported by Australian Research Council Project IH-180100002. We are grateful to the AC panel and reviewers for their comments.

References

[1] Andra Adams, Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson. Decoupling facial expressions and head motions in complex emotions. In *2015 International conference on affective computing and intelligent interaction (ACII)*, pages 274–280. IEEE, 2015. 2

[2] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29:343–351, 2016. 4

[3] Jia-Ren Chang, Yong-Sheng Chen, and Wei-Chen Chiu. Learning facial representations from the cycle-consistency of face. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9680–9689, 2021. 2, 3, 5, 6, 7

[4] Ya Chang, Changbo Hu, Rogerio Feris, and Matthew Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006. 2

[5] Yanan Chang and Shangfei Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20417–20426, 2022. 2

[6] 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. 1, 2, 3, 6, 7

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 6

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 2, 6

[10] Zhe Chen, Jing Zhang, and Dacheng Tao. Progressive lidar adaptation for road detection. *IEEE/CAA Journal of Automatica Sinica*, 6(3):693–702, 2019. 2, 3

[11] Zhe Chen, Jing Zhang, and Dacheng Tao. Recurrent glimpse-based decoder for detection with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5260–5269, June 2022. 2

[12] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 5

[13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005. 6

[14] Samyak Datta, Gaurav Sharma, and CV Jawahar. Unsupervised learning of face representations. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (2018)*, pages 135–142. IEEE, 2018. 6, 7

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 7

[16] Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. Understanding and improving lexical choice in non-autoregressive translation. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[17] Liang Ding, Longyue Wang, and Dacheng Tao. Self-attention with cross-lingual position representation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

[18] 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. In *International Conference on Learning Representations*. 2

[19] Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *British Machine Vision Conference (BMVC)*, page 104, 2019. 6

[20] Justin A Gamble and Jingwei Huang. Convolutional neural network for human activity recognition and identification. In *2020 IEEE International Systems Conference (SysCon)*, pages 1–7. IEEE, 2020. 1

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4

[22] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun

Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, (64):59–63, 2015. 5

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 6

[24] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020. 4

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 2, 6

[26] 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. 1, 2, 3, 6, 7

[27] 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. 2, 3

[28] Wei Hu, Yangyu Huang, Fan Zhang, Ruirui Li, Wei Li, and Guodong Yuan. Seqface: make full use of sequence information for face recognition. *arXiv preprint arXiv:1803.06524*, 2018. 2

[29] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 5

[30] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv preprint arXiv:1711.04598*, 2017. 1

[31] A Sophia Koepke, Olivia Wiles, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *British Machine Vision Conference (BMVC)*, page 302, 2018. 2, 6, 7

[32] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 5

[33] 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. 7

[34] Yong Li, Jiabei Zeng, and Shiguang Shan. Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):302–317, 2020. 2

[35] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019. 2, 3, 6, 7

[36] Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, 2017. 4

[37] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019. 5

[38] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 7

[39] Yuanyuan Liu, Wei Dai, Fang Fang, Yongquan Chen, Rui Huang, Run Wang, and Bo Wan. Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition. *Information Sciences*, 578:195–213, 2021. 6

[40] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023. 2

[41] Liupei Lu, Leili Tavabi, and Mohammad Soleymani. Self-supervised learning for facial action unit recognition through temporal consistency. In *British Machine Vision Conference (BMVC)*, 2020. 6

[42] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database.

IEEE Transactions on Affective Computing, 4(2):151–160, 2013. 6

[43] UNM MRN. Learning deep representations by mutual in-formation estimation and maximization. *stat*, 1050:22, 2019. 1

[44] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 5

[45] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. 6, 7

[46] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 2

[47] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 7

[48] Shuvendu Roy and Ali Etemad. Self-supervised contrastive learning of multi-view facial expressions. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 253–257, 2021. 2, 3

[49] Sebastian Ruder and Barbara Plank. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, 2018. 4

[50] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013. 6

[51] Atanu Samanta and Tanaya Guha. On the role of head motion in affective expression. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2886–2890. IEEE, 2017. 2

[52] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018. 7

[53] Yuxuan Shu, Xiao Gu, Guang-Zhong Yang, and Benny Lo. Revisiting self-supervised contrastive learning for facial expression recognition. In *BMVC*, 2022. 2, 3

[54] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–665, 2018. 7

[55] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, May 2015. 4

[56] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 1, 2

[57] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 2

[58] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 2

[59] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 6

[60] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 7

[61] Kaili Zhao, Wen-Sheng Chu, and Aleix M Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 2090–2099, 2018. 1

[62] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016. 1, 7

[63] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *British Machine Vision Conference (BMVC)*, page 317, 2018. 6

[64] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5:7, 2018. 5

[65] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 2, 3

[66] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10306–10315, 2021. 2, 6

[67] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 6