

# Revisiting Self-Supervised Contrastive Learning for Facial Expression Recognition

Yuxuan Shu<sup>1</sup>

y.shu21@imperial.ac.uk

Xiao Gu<sup>1</sup>

xiao.gu17@imperial.ac.uk

Guang-Zhong Yang<sup>2</sup>

gzyang@sjtu.edu.cn

Benny Lo<sup>1</sup>

benny.lo@imperial.ac.uk

<sup>1</sup> The Hamlyn Centre

Imperial College London

London SW7 2AZ, UK

<sup>2</sup> The Institute of Medical Robotics

Shanghai Jiao Tong University

Shanghai 200240, China

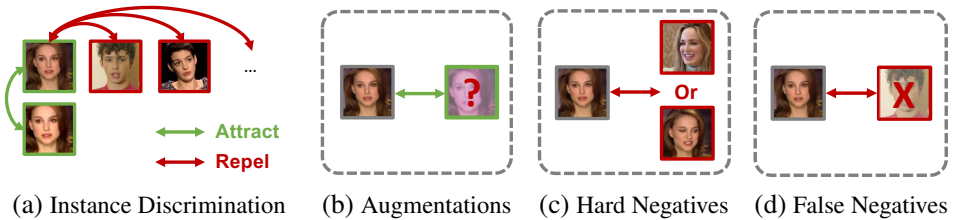
## Abstract

The success of most advanced facial expression recognition works relies heavily on large-scale annotated datasets. However, it poses great challenges in acquiring clean and consistent annotations for facial expression datasets. On the other hand, self-supervised contrastive learning has gained great popularity due to its simple yet effective instance discrimination training strategy, which can potentially circumvent the annotation issue. Nevertheless, there remain inherent disadvantages of instance-level discrimination, which are even more challenging when faced with complicated facial representations. In this paper, we revisit the use of self-supervised contrastive learning and explore three core strategies to enforce expression-specific representations and to minimize the interference from other facial attributes, such as identity and face styling. Experimental results show that our proposed method outperforms the current state-of-the-art self-supervised learning methods, in terms of both categorical and dimensional facial expression recognition tasks. Our project page: <https://claudiashu.github.io/SSLFER>.

## 1 Introduction

Facial expression recognition (FER) plays an important role in a series of applications, ranging from human-computer interaction [9], social robotics [8], to mental health monitoring [12]. Recently, considerable research efforts in computer vision have been dedicated to developing systems that can understand facial expressions from facial images automatically, including basic emotion categories in the categorical level [10], as well as valence and arousal in the dimensional level [6].

One of the reasons leading to the success of most FER systems is the availability of large-scale annotated facial datasets [21, 28]. However, the curation of such datasets poses several challenges over the course of acquiring labels. In fact, annotating real-world facial images requires significant amount of time/efforts and high-level expertise [21]. Even worse, different levels of expertise may induce annotation bias, leading to inconsistent or noisy labels [24], and the category distribution is usually imbalanced [14] such that the performance



**Figure 1: Illustration of the inherent issue and potential solutions when applying self-supervised instance-level contrastive learning for facial expression recognition.** (a) The conventional paradigm for self-supervised contrastive learning aims to pull close image representation from the same instance together, where emotion irrelevant features, e.g. human identities and hair styles, would be easily learned as shortcuts. In this paper, we propose three core directions to tackle this. (b) **Explore effective data augmentations**, Section 3.2. (c) **Increase hard negative pairs**, Section 3.3. (d) **Reduce false negative pairs**, Section 3.4. Pictures are selected from VoxCeleb [7, 29, 30].

of direct supervised learning is limited. Faced with these issues, there is a pressing need for minimizing the reliance of annotated data, yet achieving satisfactory FER performance, and self-supervised learning can potentially be the solution.

Self-supervised learning (SSL) is an emerging research line in deep learning, which aims to derive semantically meaningful representations by self-designed proxy tasks [25]. Among existing solutions, contrastive based SSL has demonstrated reasonable performance by instance-level discrimination [6, 13, 15]. It functions by maximizing the similarity, in the representation level, of the same image under different augmentations/views (positive pair), whereas minimizing those of different instances (negative pair). Despite its success on several vision tasks, there remain inherent issues of such instance discrimination strategy [6, 15], which become even more apparent when being applied on FER.

As shown in Figure 1, given a batch of samples drawn from different subjects, what would be learned when promoting the similarity of images under different augmentations? In fact, facial images are complicated, composed of multiple factors/attributes including expressions, head poses, identity, makeup, hair styles, etc. [9]. It is hardly possible to regulate the network to learn expression-relevant features only, without knowing the actual expression label. We hereby revisit self-supervised contrastive learning, and argue that there are three main promising solutions for better facial expression-related representation learning when performing instance discrimination.

**1. Exploring More Effective Augmentations for Positive Pairs.** Stronger augmentations have been proven to be able to facilitate better representation learning [24]. Conventional spatial augmentation operations utilized in exiting visual image recognition works [6, 15], like crop/resize and horizontal flipping, may not be sufficient for extracting robust expression related information.

**2. Increase Hard Negative Pairs.** Hard negative pairs represent images with different downstream labels, yet not easy to be differentiated [15]. Increasing the number of hard negative pairs during training is able to force the network to learn task related features. However, without knowing the actual emotion label in self-supervised settings, it is difficult to determine which is hard and which is negative. It remains open to develop an effective strategy to pick up the real hard negative pairs to facilitate training.

**3. Reduce False Negative Pairs.** Furthermore, given a large batch of data, there would

be quite a few samples having similar expressions, yet treated as negative pairs to be pushed away. The existence of such false negative pairs would further limit the performance of self-supervised contrastive learning [10]. However, it is challenging to reduce false negative pairs without knowing the actual labels.

Therefore in this paper, in line with these three core directions, we propose an effective self-supervised contrastive learning framework for FER. Based on expression related spatio-temporal augmentations and the region-level similarity in similar expressions, we complementary intergrate several novel strategies to mitigate the above three issues in FER. The proposed method was evaluated on two different FER tasks.

## 2 Related Works

**Facial Expression Learning.** Thus far, considerable research efforts have been devoted to facial expression recognition, in terms of advanced computational models that are able to extract discriminative FER features [11, 12, 13]. Despite these advances, most of these solutions rely heavily on annotated training data, which would fail when faced with insufficient high-quality and consistent labels. A few recent works have proposed the use of self-supervised learning for expression [14, 15]. They are heavily dependent on specific datasets, whereas multi-view images [16] or additional modalities [17] are required. It remains an open research question on how to design effective SSL tasks, to effectively extract expression-specific information.

**Facial Attribute Decomposition.** Another line of research is targeted at decomposing different attributes from facial representations. This is important for learning robust expression-related features, since it reduces the spurious correlation caused by other irrelevant attributes (such as styling and identity). Existing works have attempted to perform disentanglement of identity and facial expressions [18], which however mostly are conducted in a supervised manner. On the other hand, recent studies have investigated unsupervised “Deepfake” (face swapping) approaches [19, 20, 21]. One representative work [9] proposed a cycle-consistent framework to disentangle “expression” and identity apart. However, since facial embeddings are complicated, the authors also indicated that other attributes, such as head poses, could also be learned into the expression representations. This degrades the generalization of FER, and as reported in [9], the proposed self-supervised training is still inferior to pure fully-supervised training for FER by a large margin.

**Self-Supervised Contrastive Learning.** In recent studies, self-supervised learning has been proven to have a great potential in learning good visual representations [22] by either generative or contrastive, or the combinations [23]. Contrastive learning is emerging as a simple yet effective self-supervised manner, based on instance discrimination [8, 13, 15, 24]. However, its success relies on the assumption that the training batch are i.i.d (independently and identically distributed) sampled [25], which however is rarely the case for facial expression datasets [26], since the facial expressions are complicated with many other irrelevant factors, such as head poses and identity-related features. It is of paramount importance to avoid the shortcuts caused by these irrelevant factors, yet difficult without knowing the labels.

Approaches could be focused on either the instances that are going to be attracted (related to data augmentations [27]), or to be repelled (related to hard pair selection [10, 28]); however, these have not been systematically explored in the challenging FER task yet. This leads to the questions that we want to discuss:

- In terms of image augmentation, what information should be preserved or discarded to boost the facial-expression representation learning.

- Is there any information that could act as a supervisory signal to help enhance the model performance in FER related contrastive learning.

## 3 Methodology

### 3.1 Notations and Problem Definition

Contrastive learning works by clustering the positive pairs (usually generated with different augmentations on the same image or aligned modality), whereas pushing away the negatives (usually generated from other images). Given a batch of data  $X = \{x_1, x_2, \dots, x_N\}$ , the *InfoNCE Loss* for self-supervised instance discrimination is formulated as in Equation (1).

$$\mathcal{L}_N = \mathbb{E}_X \left[ -\log \frac{e^{f_{sim}(x_i, x_j)/\tau}}{\sum_{k=1}^N \mathbb{1}_{[k \neq i, j]} e^{f_{sim}(x_i, x_k)/\tau}} \right], \quad (1)$$

where  $N$  is the number of images in a batch,  $\tau$  is a temperature constant, and  $f_{sim}$  is the similarity matrix of feature representations, which is cosine similarity by default. For each anchor image  $x_i$ , the image  $x_j$  (different augmentations of  $x_i$ ), is considered as positive. The objective of this loss is to minimize the distance between positive pairs, while maximizing that between other negative image pairs.

However, for self-supervised FER, it is challenging to regulate the network to only learn expression-related representations without the interference of other attributes, if no prior knowledge is given. To tackle this, we present our solutions as below.

### 3.2 Positives with Same Expression

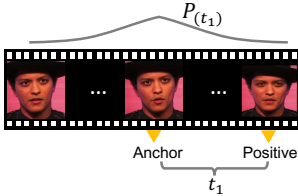


Figure 2: **TimeAug**: We sample the video sequence (Voxceleb1 [29]) along the time domain to perform time augmentation.

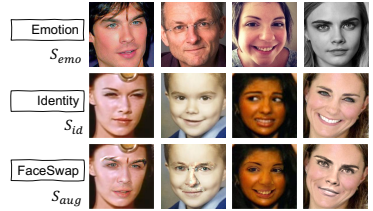


Figure 3: **FaceSwap**: Emotion is transferred to another face with different identities to synthesize positive pairs. (Pictures selected from AffectNet [28] for demonstration)

First of all, to explore effective augmentations, we developed stronger augmentation strategies mainly based on two assumptions: 1) Expressions of human beings tend to change slightly within a short time interval. 2) The facial appearance indicates more identity-related information compared to facial landmark structural characteristics.

#### 3.2.1 Temporal Augmentation

We adopt temporal augmentation (**TimeAug**) by adding temporal shifts to positive pair generation. It is an intuitive concept that two samples sampled from the same video possess higher similarity with smaller time intervals. Based on this observation, we assume that within a short time period, facial expressions would not vary too much, and the shorter the time interval, the more similar their expressions would be. In practice, inspired by [62], the sampled time interval  $t_1$  follows a downscale distribution over  $[0, T_1]$  (Figure 2), where  $T_1$

represents the maximum time interval that is considered as positives (empirically 1 second) and  $P(t_1)$  stands for the probability of sampling  $t_1$ .

### 3.2.2 FaceSwap

Whilst adding **TimeAug** to the training procedure enhances the capability of extracting expression-related information, it remains unresolved how to pull close the images of different persons which are supposed to act as positives (i.e. the same expression). To facilitate this, we resort to a simple yet effective strategy **FaceSwap** to generate fake faces that exhibits a similar expression as shown in Figure 3, built on the assumption that facial appearance represents more identity-related information rather than expressions compared to facial landmark structures. The procedure is illustrated as below:

1. Given a positive sample  $S_{emo}$  (provides the same expression as the Anchor image) and an image  $S_{id}$  that is randomly selected from the whole training set (provides a different identity from the Anchor image), extract the landmarks of the two images.
2. Align the two images with colour correction and topology transformation according to the facial landmarks.
3. Replace the region within the landmark in  $S_{id}$  with  $S_{emo}$ .

To the best of our knowledge, this is the first work that adopts face swapping strategy into contrastive learning as an augmentation strategy. It should be noted that this procedure is similar to another line of facial works, “Deepfake” [23, 51, 56]. Instead of utilizing those deep-learning based face-swapping methods, our strategy has already demonstrated its efficacy and is far easier and more computationally efficient. In practice, we performed on 50% of the positive images, because our proposed simple operation might generate weird faces when there are significant facial differences, and structural information cannot fully represent facial expressions.

### 3.3 Negatives with Same Identity

To further avoid the shortcuts caused by identity-related information during instance discrimination, we propose **HardNeg**. On top of the original negative pairs from other identities, we sample images from the same subject but with large time interval as “hard negative”. Given two facial images from the same identity, the representation will least prefer identity-related information if we are going to push away them. Hard negative pairs with  $t_2$  interval are randomly sampled from  $T_2$  onwards, where  $T_2$  is the minimum time interval that is considered as hard negative (empirically as 3 seconds).

Thus far, the whole training pipeline is shown in Figure 4, with both effective augmentations for positive pair generation and sampling policy for hard negative pair selection being incorporated into. However, the negative, including the hard negatives we select, could potentially be false negatives. The urgent issue on false negatives is still unresolved.

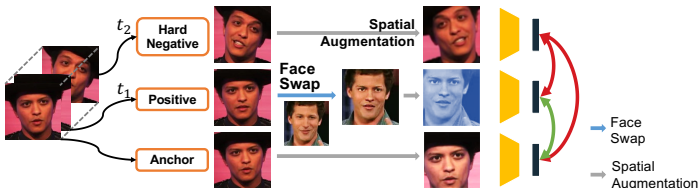


Figure 4: **Illustration of the positive and hard negative pair generation.** Positive images are chosen with random **TimeAug** and augmented randomly with **FaceSwap**. Meanwhile, hard negative pairs are selected from the images with the same identity but large time intervals to avoid identity-related information to be learned. (Demonstrate with Voxceleb1 [29].)

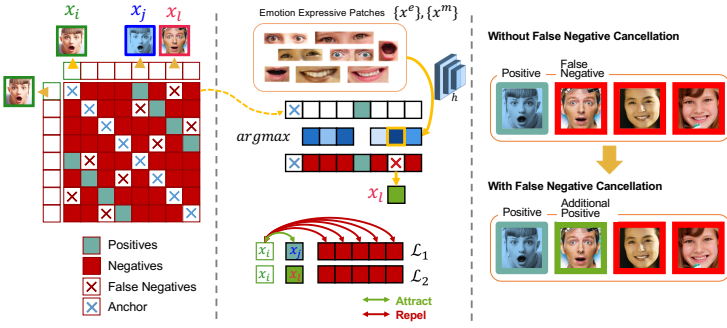


Figure 5: **Illustration of false negative cancellation.** Eyes and mouths are cropped as expressive indicators, with features being extracted and concatenated for similarity matrix calculation. We pick the sample with the highest similarity to the anchor, considering it as a false negative. Subsequently, we calculate the loss based on the original positive target  $\mathcal{L}_1$  as well as the picked false negative  $\mathcal{L}_2$ . Pictures are selected from AffectNet [28].

### 3.4 False Negatives Cancellation

In FER, false negatives are the images that are treated as negatives, yet they actually present similar expressions to the anchor image. False negatives are undesirable since they impair the training process by discarding the semantic information of the same expression, especially when there are only a small number of classes when training with a large batch size. Unfortunately, this is difficult to avoid without knowing the actual labels. Thus we propose **MaskFN** to help minimizing the negative effects caused by false negatives during training, as shown in Figure 5. Here we consider the facial expressions as discrete emotions [24].

In fact, some landmark areas, like eyes and mouth, usually contain more explicit expression signs than the other facial areas. For instance, the emotion “Happy” usually contains a smile with the corners of the mouth turning up, while “Fear” usually comes with wild open eyes and mouth. Robert and Adam [54] studied the low-level mouth feature-based emotion classification, and have proved that mouth could act as an expressive indicator of emotion.

On the basis of the above observation, we assume that:

- High-level features of eyes and mouths have higher similarity for those facial images with similar emotions.
- Those recognized as false negatives can be set as positives of the anchor image.

Given the assumptions and the fact that the batch size is much larger than the class numbers, in a mini-batch, the pair with most similar eye or mouth regions are of high probability to be false negatives. Therefore, we propose a novel strategy to identify the potential false negative by utilizing the characteristics of eyes and mouth, implemented as follows,

1. For each image  $x$  in  $X$ , extract the landmarks of the face, and crop the region of eyes  $x^e$  and mouth  $x^m$  accordingly.
2. Obtain  $z^e, z^m = h(x^e), h(x^m)$ , where  $h$  is the feature encoder (*ResNet18 pretrained on ImageNet*) and  $z^e, z^m$  are the globally averaged features.
3. Concatenate  $z^e, z^m$  into  $z^{cat}$  and calculate  $f_{sim}$  of feature  $z^{cat}$  with cosine similarity.
4. Set  $N_{FN}$  ( $N_{FN} \ll N$ ) images with the highest similarity (that is originally neither positive nor hard negative) as additional positives.

The proposed false negative cancellation pipeline is shown in Figure 5. Without loss of generality and for clarity, we only discuss picking up one false negative in **MaskFN** for each anchor by default in this paper, if not specified. The contrastive loss is calculated as shown in Equation (2), where  $x_l$  stands for the additional positive image with respect to anchor image



$x_i$ .  $\mathcal{L}_2$  is assigned with less weight (empirically set as  $\frac{1}{2}$ ), since the selected false negative images may not be the true positive pairs as we only consider the similarity of mouth and eyes as criteria. The total loss is formulated as  $\mathcal{L}_1 + \frac{1}{2}\mathcal{L}_2$ .

$$\mathcal{L}_1 = \mathbb{E}_X \left[ -\log \frac{e^{f_{sim}(x_i, x_j)/\tau}}{\sum_{k=1}^N \mathbb{1}_{[k \neq i, j, i]} e^{f_{sim}(x_i, x_k)/\tau}} \right], \quad \mathcal{L}_2 = \mathbb{E}_X \left[ -\log \frac{e^{f_{sim}(x_i, x_i)/\tau}}{\sum_{k=1}^N \mathbb{1}_{[k \neq i, j, i]} e^{f_{sim}(x_i, x_k)/\tau}} \right]. \quad (2)$$

## 4 Experiments

### 4.1 Experiment Setup

**Self-supervised pre-training.** All self-supervised learning network was trained on NVIDIA GeForce RTX 3090 from scratch. ResNet50 was used as the backbone network for feature extraction. Following SimCLR [6], two MLP (Multi-Layer-Perceptron) layers were added with a default 128 feature dimension. The training batch size was 256 with a initialised  $3e-4$  learning rate and  $1e-4$  weight decay. We optimized the network using Adam optimizer except for MoCo-v2 [6], BYOL [13] optimized with SGD. We applied a cosine annealing schedule from 10 epochs onwards. The temperature  $\tau$  was set to 0.07 in the InfoNCE loss for all experiments. During this stage, we saved the checkpoints every 5 epochs when its instance discrimination top-1 accuracy reached 60% for downstream task evaluation.

**Downstream evaluation.** We evaluated the pretraining performance by freezing and fine-tuning the encoder with three MLP layers on top. For fair comparison, we kept the same set of hyper-parameters across all models for each downstream task, with a batch size of 64. Training was conducted using an Adam optimizer over 20 iterations for AffectNet dataset and 300 iterations for FER2013 dataset with a weight decay of  $5e-4$ . Initially, the learning rate was set to  $1e-4$  and decayed with cosine annealing learning rate scheduler.

**Data augmentation.** The adopted spatial augmentation for both contrastive learning and downstream evaluation include landmark-based masking (eyes or mouth), resizing (128, 128), random crop (112, 112), horizontal flipping, Gaussian blurring, colour jittering, and random grayscale. We applied **TimeAug**, **FaceSwap** only in self-supervised pre-training.

**Landmarks.** To facilitate **FaceSwap** in Section 3.2.2, we used dlib toolkit [20] to detect landmarks and the bounding box of eyes and mouth. The extracted landmark contains 68 points, including jaw, left eyebrow, right eyebrow, left eye, right eye, nose and mouth.

### 4.2 Datasets

Three datasets were used in our work, namely **VoxCeleb1**, **AffectNet** and **FER2013**.

The **VoxCeleb1** [29] dataset is a large-scale facial dataset, which includes videos that were extracted from YouTube of 1251 celebrities with different ages and of different ethnicity. It provides image sequences yet no emotion related annotations.

The **AffectNet** [28] dataset is the largest facial expression annotated (classification and regression) in the wild dataset which contains more than 1,000,000 facial images from different genders and races. We used all 8 classes of emotions (neutral, happy, angry, sad, fear, surprise, disgust, contempt) for classification on the publicly available training and validation splits, with the same settings as in [40].

The **FER2013** [10] dataset is a widely used dataset in grayscale which consists of 28,709 training images. We reported the results on the test split, with the best performance model on the validation split, following the same settings as [27, 39].

The proposed method was trained on the **VoxCeleb1** [29]. After training, the performance of the learnt representation was validated on **AffectNet** [28] and **FER2013** [10] with two downstream tasks, namely Emotion Classification and Valence & Arousal Recognition.

### 4.3 Downstream Tasks

**Emotion Classification:** Macro  $F_1$  score,  $F_1 = \frac{1}{n} \sum_{i=1}^n \mathcal{F}_1^i$ , was used for evaluating categorical expression classification. Since the training set may suffer from imbalance, we adopted the state-of-the-art Balanced Softmax Cross Entropy Loss [53] for training.

**Valence & Arousal Recognition:** We also reported results of Valence & Arousal recognition, where two common metrics were used for evaluating both the error and correlation.

They are Root Mean Square Error  $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2}$  and Concordance Correlation Coefficient (CCC) [42]  $\rho_c = \frac{2\rho\sigma_{\hat{\theta}}\sigma_{\theta}}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2 + (\mu_{\hat{\theta}} - \mu_{\theta})^2}$ , respectively. We applied CCC loss for regression optimization following [46].

### 4.4 Quantitative Results

Pretraining Methods		Dataset		Freeze-AffectNet		Finetune-AffectNet						Finetune-FER2013		
				EXPR		EXPR		Valence		Arousal		EXPR		
				F1↑	Acc↑	F1↑	Acc↑	CCC↑	RMSE↓	CCC↑	RMSE↓	Acc↑		
Supervised		ImageNet	-	-	56.7%	56.6%	0.563	0.462	0.480	0.376	69.12%			
BYOL [43]		VoxCeleb1	34.1%	37.2%	56.3%	56.4%	0.560	0.460	0.462	0.386	68.98%			
MoCo-v2 [6]		VoxCeleb1	38.0%	38.1%	56.8%	56.8%	0.570	0.454	0.486	0.378	69.13%			
SimCLR [8]		VoxCeleb1	53.4%	53.5%	57.5%	57.7%	0.594	0.431	0.451	0.387	68.45%			
CycleFace [9]		VoxCeleb1.2	-	-	48.8%	49.7%	0.534	0.492	0.436	0.383	69.86%			
Ours	TimeAug	HardNeg	FaceSwap	MaskFN	VoxCeleb1	53.9%	54.1%	57.8%	57.9%	0.583	0.448	0.500	0.374	69.32%
	a ✓				VoxCeleb1	55.4%	55.5%	58.1%	58.3%	0.594	0.437	0.500	0.373	69.60%
	b ✓	✓			VoxCeleb1	54.2%	54.2%	58.3%	58.4%	0.542	0.463	0.508	0.368	69.15%
	c ✓	✓	CutMix [43]		VoxCeleb1	55.6%	55.8%	58.6%	58.7%	0.568	0.444	0.502	0.369	70.07%
	d ✓		✓	✓	VoxCeleb1	56.0%	56.0%	58.8%	58.9%	<b>0.601</b>	<b>0.429</b>	<b>0.514</b>	<b>0.367</b>	70.47%
	e ✓	✓	✓		VoxCeleb1	<b>57.1%</b>	<b>57.1%</b>	58.9%	58.9%	0.578	0.448	0.493	0.370	70.21%
	f ✓	✓	✓	✓	VoxCeleb1	56.4%	56.4%	<b>59.3%</b>	<b>59.3%</b>	0.595	0.435	0.502	0.372	<b>71.66%</b>

Table 1: Comparison of FER performance on AffectNet and FER2013, in terms of categorical expression classification, and valence & arousal recognition. The results were based on linear evaluation (Freeze) and fine-tuning (FineTune).

In this section, quantitative results are presented. For comparison, we also implemented state-of-the-art self-supervised learning methods (SimCLR [8], MoCo-v2 [6], BYOL [43]) pretrained on **Voxceleb1** from scratch, as well as utilizing the model weights pretrained on ImageNet [8], as shown in Table 1. In the last four rows of Table 1, we reported the performance of the facial representation learning by adding each strategy separately or jointly. As noted, our proposed strategies significantly outperform both the ImageNet-pretrained model as well as other self-supervised methods, on all tasks.

In addition, we compared with one state-of-the-art unsupervised deepfake work, CycleFace [9], which encompasses an identity-irrelevant branch to encode facial expressions. We finetuned its emotion extraction branch and it yielded inferior performance across all metrics (beige row). Our conjecture is that not only expression, but other attributes that are irrelevant to the expression, such as head poses, are encoded to the expression branch as well, which was also raised by the authors [9].

To evaluate the efficacy of **FaceSwap**, we also trained with CutMix [43] as shown in Table 1 row (c) by only cutting and mixing the centre patch of the images. Comparing it with Ours as shown in row (e), our superior performance demonstrates that the proposed **FaceSwap** leads to better FER representations.

### 4.5 Qualitative Results

We visualised the saliency of contributing features with Grad-CAM [58] across different persons with different emotions. As shown in Figure 6 (a), the model trained with Sim-



CLR mostly unsatisfactorily attend on emotion-irrelevant areas. By contrast, our proposed **FaceSwap** and **MaskFN** focus more on eyes and mouths, indicating that the proposed strategies are able to regulate the network to focus more on the regions that are more expressive to emotions. Moreover, the focus slightly changes with different emotions. “Neutral” focuses more on eyes while “Happy” and “Fear” focus more on eyes and mouths. Additionally, we also visualised the saliency maps of the same subject across time in Figure 6 (b). It can be observed that with the same subject, the focus of our method slightly changes as the expression changes, indicating that our network is able to attend on expression-related regions.

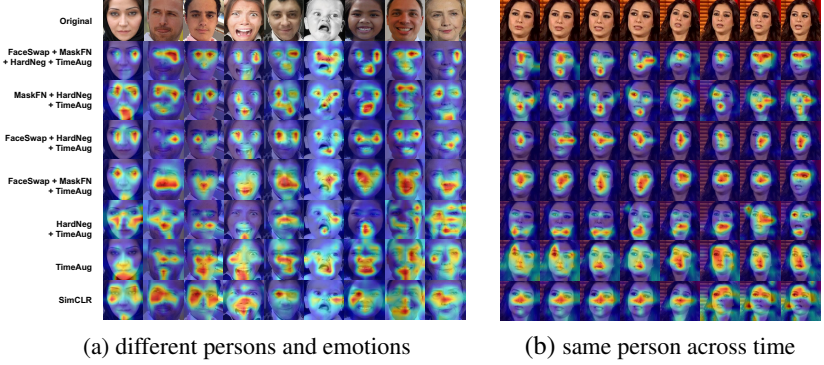


Figure 6: Saliency maps using Grad-CAM [38] comparing different strategies. Pictures are selected from AffectNet [28] (a) and Voxceleb [7, 29, 30] (b).

## 4.6 Additional Experiments

**Influence on different strategies:** Table 1 shows the effectiveness of each module. It can be observed that they are effective for the emotion classification task, while **MaskFN** leads to poorer result on the Valence & Arousal Recognition. One possible reason is that our assumption for false negative cancellation is based on concrete categorical emotions, whereas those additional “positive” by false negative cancellation may not be that “positive” for Valence/Arousal regression. Please see further discussions in Section 4.7.

SSL Methods		FR-KNN ↓
BYOL [16]		64.5%
MoCo [9]		63.8%
SimCLR [9]		63.5%
Ours	TimeAug	
	HardNeg	
	FaceSwap	
	MaskFN	
	a	64.1%
	b	63.9%
	c	62.9%
	e	59.5%
f	✓	58.2%
	✓	56.5%
	g	56.5%

Table 2: Face recognition (FR) performance on the LFW dataset using KNN.

Methods	Epochs	Top1-Acc	FR-KNN	FER-Finetune	
				F1↑	Acc↑
SimCLR	5	85.7%	55.2%	52.5%	52.7%
	10	90.8%	61.7%	55.0%	55.1%
	30	95.3%	62.6%	56.8%	56.9%
	50	97.1%	61.6%	57.6%	57.7%
	150	99.0%	61.4%	55.4%	55.5%
Ours	150	65.4%	59.5%	59.1%	59.1%
	180	75.6%	58.0%	59.2%	59.3%
	210	85.1%	56.5%	59.3%	59.3%
	250	90.0%	58.7%	58.3%	58.5%

Table 3: Results of different training stages.

**Facial recognition performance:** We further evaluated our proposed method with face recognition (FR) on the LFW dataset [16]. We used the pre-trained network to extract the facial representations and performed verification based on KNN, the same settings as Cycle-Face [9]. As shown in Table 2, the performance on FR degraded with our proposed strategy, in line with our objective of removing identity-related information to avoid shortcuts in FER.

**Influence on training stages:** Table 3 reports the performance of the model on classification when training at different stages, where the second column shows the training epochs and the third column indicates top-1 accuracy of instance discrimination. Self-supervised training might get overfitted after certain stages on facial images, deteriorating the downstream FER tasks. This can be observed in both SimCLR and Ours, indicating that the model is likely to learn identity-related information if overtrained.

Methods	EXPR		Valence		Arousal	
	$F_1 \uparrow$	Acc $\uparrow$	CCC $\uparrow$	RMSE $\downarrow$	CCC $\uparrow$	RMSE $\downarrow$
MaskFN(1)	59.3%	59.3%	<b>0.595</b>	<b>0.435</b>	<b>0.502</b>	<b>0.372</b>
MaskFN(2)	<b>59.7%</b>	<b>59.7%</b>	0.583	0.444	0.484	0.378

Table 4: Influence on having different numbers (as in brackets) of false negatives. (With TimeAug, HardNeg and FaceSwap)

FR-KNN	FER-Linear	
	F1	Acc
45.5%	51.0%	51.1%

Table 5: Eye-Mouth Descriptor.

**Influence on having different number of false negatives:** As shown in Table 4, introducing an additional false negative improves the classification performance by 0.5% yet impairs the performance on Valence/Arousal regression; this further validating our hypothesis that **MaskFN** selects false positives that are more “positive” in terms of categorical emotions.

**Validation of mouth-eye descriptors:** We provide further experiments to validate that the mouth and eye have more emotional-related information and could act as expressive indicators. We used the feature of mouth and eye extracted from fixed ResNet18 (pretrained on ImageNet), which is identical to what we used in MaskFN. Subsequently, linear probing was performed on AffectNet for FER, and KNN on LFW for FR. The result shows that it has superior performance in FER, yet inferior in FR as shown in Table 5.

## 4.7 Limitations

Albeit the improvement **MaskFN** has brought about to expression classification task, the performance of learned representation on regression task was noticeably decreased. This may probably result from the fact that our **MaskFN** strategy was built on the assumption of categorical expressions where similar emotions share similar structural characteristics in eyes and mouth. However, this may not apply to the dimensional expression analysis. Same categories may exhibit different ranges of Valence & Arousal and pushing together these samples may in turn degrade the sensitivity of VA regression. This highlights that a task-specific network design is necessary for reaching a satisfactory downstream result.

## 5 Discussions and Conclusions

Facial representations are complicated, composed of multiple attributes. Conventional contrastive based self-supervised learning fails to learn expression-specific representations. In this paper, we revisited the use of self-supervised contrastive learning, and proposed three complementary novel strategies to regulate the network to lean towards emotion related information. In particular, based on emotion related properties in facial images, we explored effective augmentations, hard negative pair sampling manner, as well as false negative cancellation strategy. The experimental results have shown that our self-supervised training strategies outperform the state-of-the-art methods on downstream FER tasks, including both categorical expression classification and dimensional valence&arousal regression.

## References

- [1] Challenges in representation learning: Facial expression recognition challenge, 2013.
- [2] Baron-Cohen and Tead. *Mind reading: The interactive guide to emotion*. Jessica Kingsley Publishers, 2003.
- [3] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. Social robotics. *Springer handbook of robotics*, pages 1935–1972, 2016.
- [4] 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.
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [9] Alan Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. *Human-computer interaction*. Pearson Education, 2004.
- [10] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2), 1971.
- [11] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021.
- [12] Zixiang Fei, Erfu Yang, David Day-Uei Li, Stephen Butler, Winifred Ijomah, Xia Li, and Huiyu Zhou. Deep convolution network based emotion analysis towards mental health care. *Neurocomputing*, 388:212–227, 2020.
- [13] 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.
- [14] Xiao Gu, Yao Guo, Zeju Li, Jianing Qiu, Qi Dou, Yuxuan Liu, Benny Lo, and Guang-Zhong Yang. Tackling long-tailed category distribution under domain shifts. In *ECCV*, 2022.
- [15] 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.

- [16] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008.
- [17] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2785–2795, 2022.
- [18] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.
- [19] Aparna Khare, Srinivas Parthasarathy, and Shiva Sundaram. Self-supervised learning with cross-modal transformers for emotion recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 381–388. IEEE, 2021.
- [20] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [21] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022.
- [22] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.
- [23] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping, 2019. URL <https://arxiv.org/abs/1912.13457>.
- [24] Shimin Li, Hang Yan, and Xipeng Qiu. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11002–11010, 2022.
- [25] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [26] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggao Zhang, Chuanhe Liu, and Qin Jin. Valence and arousal estimation based on multi-modal temporal-aware features for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2345–2352, 2022.
- [27] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9):3046, 2021.
- [28] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

- [29] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.
- [30] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.
- [31] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepface-lab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.
- [32] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.
- [33] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.
- [34] Staniucha Robert and Wojciechowski Adam. Mouth features extraction for emotion classification. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1685–1692. IEEE, 2016.
- [35] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020.
- [36] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.
- [37] 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.
- [38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [39] Adrian Vulpe-Grigorași and Ovidiu Grigore. Convolutional neural network hyperparameters optimization for facial emotion recognition. In *2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pages 1–5. IEEE, 2021.
- [40] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

- [41] Haozhe Wu, Jia Jia, Lingxi Xie, Guojun Qi, Yuanchun Shi, and Qi Tian. Cross-vae: Towards disentangling expression from identity for human faces. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4087–4091. IEEE, 2020.
- [42] 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.
- [43] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.
- [44] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018.
- [45] Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyang Shen, and Haoxin Liu. Towards unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4910–4920, 2022.
- [46] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, page 317, 2018.