Emotion-aware Multi-view Contrastive Learning for Facial Emotion Recognition

Daeha Kim and Byung Cheol Song

Inha University, Incheon, Republic of Korea
kdhht5022@gmail.com,bcsong@inha.ac.kr

Abstract. When a person recognizes another's emotion, he or she recognizes the (facial) features associated with emotional expression. So, for a machine to recognize facial emotion(s), the features related to emotional expression must be represented and described properly. However, prior arts based on label supervision not only failed to explicitly capture features related to emotional expression, but also were not interested in learning emotional representations. This paper proposes a novel approach to generate features related to emotional expression through feature transformation and to use them for emotional representation learning. Specifically, the contrast between the generated features and overall facial features is quantified through contrastive representation learning, and then facial emotions are recognized based on understanding of angle and intensity that describe the emotional representation in the polar coordinate, i.e., the Arousal-Valence space. Experimental results show that the proposed method improves the PCC/CCC performance by more than 10% compared to the runner-up method in the wild datasets and is also qualitatively better in terms of neural activation map. Code is available at https://github.com/kdhht2334/AVCE_FER.

Keywords: Facial emotion recognition, dimensional model of emotion, human-computer interaction

1 Introduction

Facial emotion (or expression) is the most distinct attention information among human non-verbal cues. Facial emotion recognition (FER) has made significant technological progress in recent years, and it has been gradually extended to various fields such as robot-assisted therapy [27] and robot navigation [2]. However, since most FER methods are based on discrete (emotion) labels, they could not take into account the intensity of emotion or capture the continuous emotional change.

Therefore, Arousal-Valence (AV)-based FER utilizing continuous labels has been studied to overcome the above-mentioned limitations of categorical FER. Here, continuous AV space is based on activation (arousal) and positiveness (valence) of emotions [18]. Psychological studies [45] showed that human visual attention is closely related to AV value(s), which suggests that AV-based FER

can imitate human's emotion recognition. Comparing with categorical FER, AV-based FER handling continuous labels can theoretically understand complex facial expressions and micro-facial expressions, and even detect hidden emotions [37,44].

However, the existing AV-based FER approaches have not yet technically dealt with the following concerns.

• *What is the key for feature learning of facial emotions?* According to Panda et al. [36] and Wei et al. [52], visual features for standard vision tasks such as classification and detection cannot be scaled up for FER-related tasks because FER should consider the pixel-level visual properties (e.g., edge, hue, illumination) as well as semantic features (e.g., landmarks). Prior arts for representation learning [14,19] learned facial emotions only through quantitative differences, so they did not provide an explicit solution for learning semantic features. Therefore, representation learning that can understand even semantic features is required.

• *How can we extract facial emotion-aware features?* In general, a human has a so-called visual perception ability that attends core regions such as eyes and mouth for FER while suppressing relatively unnecessary parts such as hair and background [16]. This fact suggests that properly extracting the features of core and non-core regions is a pre-requisite in representation learning for FER. So, it is necessary to extract emotion-aware features from the (latent) feature space that can learn semantic information [3]. However, due to the difficulty of the problem setting, AV-based FER that considers representation learning and visual perception ability simultaneously has not yet been reported as in Fig. 1.

This paper addresses the two concerns mentioned above. First, we propose a novel contrastive representation learning (CRL) mechanism and analyze the (semantic) feature relationship, i.e., emotional contrast (cf. Sec. 3.3). The proposed CRL with the similarity function performs discriminative learning based on projected features in the AV space (see the blue dotted box of Fig. 2(*b*)) [4,12]. This CRL mechanism is suitable for the FER task, since it is important to differentiate the core regions in which emotions are expressed from the non-core regions in which emotions are not expressed [53]. Note that utilizing CRL as a regularization term can improve the generalization ability of convolutional neural networks (CNNs) and have the same effect on continuous label-based tasks [26]. Therefore, the proposed CRL mechanism for regularization can also enhance the generalization ability of emotional representations in the AV space.

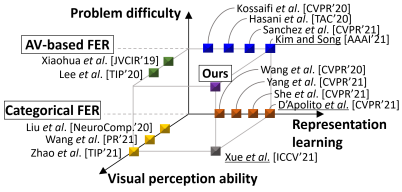

Fig. 1. Our emotion-aware representation learning is a novel method that has not been formally addressed and designed so far

Second, we propose feature transformations that generate multiple (semantic) "views" of a facial image, i.e., the facial emotion-aware features $\mathbf{z}_a$ and $\mathbf{z}_n$

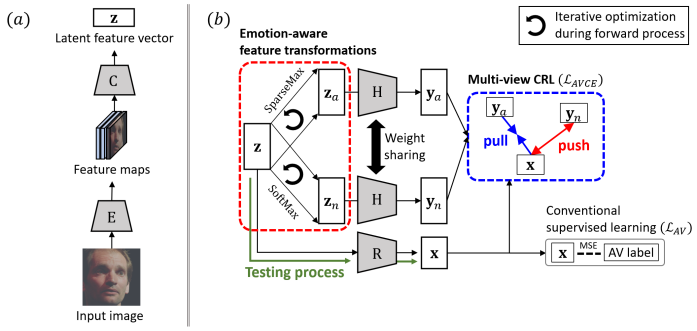

Fig. 2. (a) Input image encoding process. (b) The overall framework of the proposed method. Here, the green arrow indicates the testing process

(see the red dotted box of Fig. 2(b)). Here, for transformation that has a significant impact on the performance of CRL, SparseMax [28] and SoftMax are adopted. $\mathbf{z}_a$ from SparseMax indicates facial features that are highly correlated with facial emotion. On the other hand, since SoftMax is based on weighted aggregation of features and is detrimental to the disentanglement representation [59], $\mathbf{z}_n$ obtained from SoftMax represents an average (facial) feature or a feature contrasting with $\mathbf{z}_a$.

Therefore, the main contributions of this paper are summarized as follows:

- We succeeded in incorporating visual perception ability into representation learning for the first time in AV-based FER task. The proposed method overcame the limitations of problem setting in AV-based FER. Also, it showed better performance of more than 10% in the wild dataset than the state-of-the-art (SOTA) methods.
- The proposed feature transformations enable to focus on semantic regions that are important for emotional representation. We could observe the visual perception ability of transformed features focusing on semantic regions through activation map-based visualization.

2 Related Work

AV-based FER overview. With the advent of large-scale AV datasets [55], Hasani et al. [14] directly matched predictions using CNNs to continuous labels, i.e., ground-truths (GTs). Kossaifi et al. [23] proposed a factorized CNN architecture that achieved both computational efficiency and high performance based on low-rank tensor decomposition. However, the early methods mainly focused on quantitative differences between facial features. Only a few FER studies adopted adversarial learning capable of analyzing emotional diversity. For example, a personalized affective module based on adversarial learning of auto-encoding structure was proposed [1]. Kim and Song [19] divided image

groups according to emotional intensity and analyzed complex emotional characteristics by using adversarial learning. Also, Sanchez et al. [42] tried to encode a contextual representation for understanding the temporal dependency of facial emotions.

Representative FER approaches. Meanwhile, feature learning and various form of supervision information have been employed to overcome the limitations of discrete labels. Yang et al. [54] introduced polar coordinates on Mikel's Wheel [29] to analyze emotion polarity, type, intensity, and overcame the limitations of categorical FER through label distribution learning. However, since [54] uses simple MSE and KL divergence, it is difficult to analyze the diversity of visual emotions. Xue et al. [53] proposed a transformer [50] that learns relation-aware local representations and succeeded in learning diverse local patches of facial images for categorical FER. D'Apolito et al. [7] predicted and manipulated emotional categories through learnable emotion space instead of using hand-crafted labels. However, unlike the AV space, the discrete label-based emotion space cannot inherently handle micro-emotions. On the other hand, the proposed method that utilizes CRL based on the AV space and feature transformations can train CNNs similarly to human's FER mechanism.

Contrastive representation learning. Self-supervised learning (SSL) utilizes self-supervision which can represent hidden properties of images defined from pretext tasks. For example, Jigsaw puzzle [34] divided an image into patches and predicted shuffled patch positions. Gidaris et al. [11] predicted the angle of an image rotated by geometric transformation. CRL [4] that maximizes the agreement of self-supervisions generated by data augmentation has recently attracted attention, and CRL has been extended to multiview coding handling an arbitrary number of views [47]. Note that recent studies [40,41] have applied contrastive learning to the FER task. However, the categorical FER methods were verified only on a very limited dataset and cannot be extended to the AV space, so they are not dealt with in this paper.

3 Method

The goal of this paper is to enable CNNs to understand facial features through CRL with similarity function and feature transformations. Based on the insight and rationale derived from the latest CRL mechanism (Sec. 3.3), we propose similarity functions to describe emotional representation (Sec. 3.4), feature transformations (Sec. 3.5), and discriminative objective function (Sec. 3.6). The list below shows the nomenclature of this paper.

R, C, H	Regressor, Compressor and Projection head
$\mathbf{z}$, $\mathbf{z}_a, \mathbf{z}_n$	Latent feature and transformed features via Sparse(/Soft)Max
P_{XY}	Joint probability distribution of random variables X and Y
$P_X P_Y$	Product of marginal probability distributions
N, d	Sizes of mini-batch and latent feature

3.1 Overview

Figure 2 describes the overall framework of the proposed method. First, an encoder (E) [25,15] and a compressor (C) encode an input (facial) image and convert into a latent feature (vector), respectively. Then, the latent feature $\mathbf{z}(\in \mathbb{R}^d)$ is converted to $\mathbf{x}(\in \mathbb{R}^2)$ by a regressor (R), and then conventional supervised learning ($\mathcal{L}_{AV}$) is applied for $\mathbf{x}$ and its AV label. At the same time, $\mathbf{z}_a$ and $\mathbf{z}_n(\in \mathbb{R}^d)$ (or multiple "views") are generated by an iterative optimization-based feature transformation [10,28]. Based on a study [4] that mapping transformed features into objective function space is useful for representation learning, $\mathbf{z}_a$ and $\mathbf{z}_n$ are projected to $\mathbf{y}_a$ and $\mathbf{y}_n(\in \mathbb{R}^2)$ in the AV space through H, respectively. Then, CRL in the AV space ($\mathcal{L}_{AVCE}$) is performed through $\mathbf{x}$, $\mathbf{y}_a$, and $\mathbf{y}_n$ according to the 'push and pull' strategy. The model parameters are updated from $\mathcal{L}_{main}$ that is the summation of $\mathcal{L}_{AVCE}$ and $\mathcal{L}_{AV}$. In addition, discriminative learning ($\mathcal{L}_{dis}$) for boosting performance is applied through triplet tuple $(\mathbf{z}, \mathbf{z}_a, \mathbf{z}_n)$.

3.2 Preliminaries of CRL

Self-supervisions for CRL are designed to represent the hidden properties of an image (cf. Sec. 2), or to represent multiple views [49]. The latter aims at learning the contrast of multiple views, i.e., the gap of semantic-level information [47]. In other words, multi-view CRL injects the fact that different views $\mathbf{z}_a$ and $\mathbf{z}_n$ are contrastive each other into neural network parameters. As a result, H allows to focus on 'mouth and eyebrows', which are core regions in recognizing facial emotions (cf. Fig. 4), and helps the learning of R important for $\mathcal{L}_{AV}$ (see neural activation maps in Fig. 6). Multi-view CRL is designed through InfoNCE [35].

$$\mathcal{L}(X,Y) = \sup_{f\in\mathcal{F}} \mathbb{E}_{(\mathbf{x},\mathbf{y}_1)\sim P_{XY},\{\mathbf{y}_j\}_{j=2}^{N}\sim P_Y} \log\left(\frac{e^{f(\mathbf{x},\mathbf{y}_1)}}{\frac{1}{N}\sum_{j=1}^{N} e^{f(\mathbf{x},\mathbf{y}_j)}}\right) \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the outcomes of random variables X and Y, respectively. Note that positive pairs $(\mathbf{x}, \mathbf{y}_1)$ and negative pairs $(\mathbf{x}, \mathbf{y}_{j>1})$ are sampled from P_{XY} and $P_X P_Y$, respectively. f is a similarity function belonging to a set of real-valued functions $\mathcal{F}$.

In general, maximizing the divergence between P_{XY} and $P_X P_Y$ in Eq. (1) encourages the learned representations X and Y to have high contrast. However, Eq. (1) cannot guarantee the stability of learning [32], and it is insufficient as a theoretical basis for designing f in the AV space.

3.3 Proposed Method: AVCE

We propose the so-called AVCE suitable for learning the Contrast of Emotions in AV space while following CRL mechanism of Eq. (1) (cf. Appendix for derivation). AVCE for learning emotional representations in AV space is defined by

$$\mathcal{L}_{AVCE}(X,Y) = \sup_{f\in\mathcal{F}} \mathbb{E}_{P_{XY}} f(\mathbf{x},\mathbf{y}) - \alpha \mathbb{E}_{P_X P_Y} f(\mathbf{x},\mathbf{y}) - \frac{\beta}{2}\mathbb{E}_{P_{XY}} f^2(\mathbf{x},\mathbf{y}) - \frac{\gamma}{2}\mathbb{E}_{P_X P_Y} f^2(\mathbf{x},\mathbf{y}) \quad \text{s.t.} \quad f(\mathbf{x},\mathbf{y}) = \left(1 - \frac{\theta(\mathbf{x},\mathbf{y})}{\pi}\right) \tag{2}$$

where as many as the number of mini-batches, positive and negative pairs(or views), i.e., $(\mathbf{x},\mathbf{y}_a)$ and $(\mathbf{x},\mathbf{y}_n)$ are sampled from P_{XY} and $P_X P_Y$, respectively. α, β, and γ are relative parameters that adjust the influence between pairs. Comparing with InfoNCE (Eq. (1)), Eq. (2) without exponential or logarithmic terms guarantees learning stability, so it can converge with small variance thanks to the relative parameters acting as regularizers. Also, Eq. (2) enables mini-batch-based empirical estimation through Monte-Carlo estimation, etc. [6].

Note that f of Eq. (2) is designed as the angular similarity based on θ that can describe an emotional representation in the AV space. Here, $\theta = \cos^{-1}\left(\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\right)$ and $\|\cdot\|$ indicates L2 norm. To show that f is a function quantifying the emotional representation of pairs, we define emotional contrast (EC) and describe its property.

Definition 1. *Emotional contrast is a qualitative indicator that indicates the difference between emotions observed from two inputs (images) [38].*

If the two facial expressions look similar to each other, that is, if EC is small, then the two predicted emotions must be located close to each other in the AV space, and vice versa. The evidence that a qualitative indicator EC can be quantified through f is derived from the following Lemma.

Lemma 1. *The optimal solution of* $\mathcal{L}_{AVCE}$ *is* $f^*(\mathbf{x},\mathbf{y}) = \frac{r(\mathbf{x},\mathbf{y})-\alpha}{\beta r(\mathbf{x},\mathbf{y})+\gamma}$ *with density ratio* $r(\mathbf{x},\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}$. *Here,* $f^*(\mathbf{x},\mathbf{y})$ *is the optimal similarity that* $\mathbf{x}$ *and* $\mathbf{y}$ *can represent, and it can be obtained from the trained neural network.*

Proof. Please refer to Section A.1 in Appendix of [49].

In Lemma 1, $p(\mathbf{x},\mathbf{y})$ and $p(\mathbf{x})p(\mathbf{y})$ indicate the probability density functions of P_{XY} and $P_X P_Y$, respectively. Specifically, as the correlation of the two vectors becomes larger, the density ratio r gets larger [48]. In other words, EC and r are inversely proportional, i.e., $\text{EC} \propto \frac{1}{r}$. Also, in Lemma 1, if β is sufficiently larger than α and γ, f depends only on a constant. That is, $f \cong \frac{1}{\beta}$. So, in order to have an explicit (linear) relationship between r and the empirical estimate f, we set β to be smaller than α and γ. Then, we can approximate f as follows:

$$f(\mathbf{x},\mathbf{y}) \cong \frac{r(\mathbf{x},\mathbf{y})-\alpha}{\gamma} \propto \frac{1}{\text{EC}} \tag{3}$$

According to Eq. (3), the positive pair $(\mathbf{x},\mathbf{y}_a)$ outputs larger f than the negative pair $(\mathbf{x},\mathbf{y}_n)$ (same for r). Since $\text{EC} \propto \frac{1}{r}$, EC is also inversely proportional to f as shown in Eq. (3). Therefore, EC can be quantified with respect to θ of f.

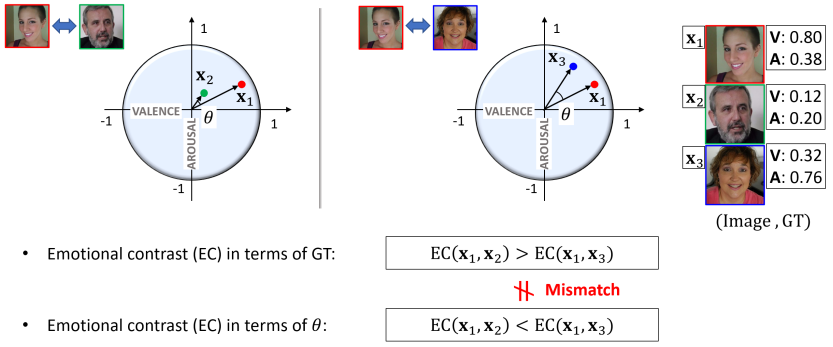

Fig. 3. A counterexample when quantifying EC using f of Eq. (2) on AffectNet dataset. Here, f only based on θ cannot properly reflect EC

Remarks. Prior arts such as [19,40] constructed contrastive samples based only on quantitative emotion labels. However, even facial images annotated with the same emotion label can express different types of emotions. Therefore, $\mathcal{L}_{AVCE}$ reflecting the visual perception ability to the contrastive loss is effective to evaluate the unseen test DB.

3.4 Similarity Function Design

Basic extension. On the other hand, EC cannot be quantified only depending on θ. Fig. 3 illustrates a counterexample. If EC is defined in terms of GT (ideal case), EC increases as the two vectors are farther apart in AV space. That is, $\mathrm{EC}(\mathbf{x}_1, \mathbf{x}_2)$ is greater than $\mathrm{EC}(\mathbf{x}_1, \mathbf{x}_3)$. However, if EC is defined only in terms of θ (cf. Eq. (2)), the opposite result is obtained based on Eq. (3). The main reason for this counterexample is that the distance between the intensity components of $\mathbf{x}_1$ and $\mathbf{x}_2$ is not considered at all. Since intensity is a factor that generally quantifies the expression level of emotion [46], it is desirable to design f considering the difference in intensity as well as the directional difference between the two vectors, i.e., angular similarity. Therefore, we redefine f of Eq. (2) as follows:

$$f(\mathbf{x}, \mathbf{y}) = \left(1 - \frac{\theta(\mathbf{x}, \mathbf{y})}{\pi}\right) + \mu\left(1 - |\|\mathbf{x}\| - \|\mathbf{y}\||\right) \tag{4}$$

where μ is a balance factor and $|\cdot|$ outputs the absolute value of the input.

Different ways to represent emotional contrast. $f_1(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\|\|\mathbf{y}\| \cos\theta$ can be an alternative to Eq. (4). However, f in Eq. (4) can consider angle and intensity independently of each other, whereas the two components are entangled in f_1. Therefore, f is more advantageous than f_1 in dealing with multicultural cases [46] in which the influences of angle and intensity are expressed differently in the AV space. Meanwhile, facial emotions tend to be grouped mainly by the valence polarity [54]. So, emotion polarity can be added as follows:

$$f_2(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) + \mathbf{1}_{pol}(\mathbf{x}, \mathbf{y}) \tag{5}$$

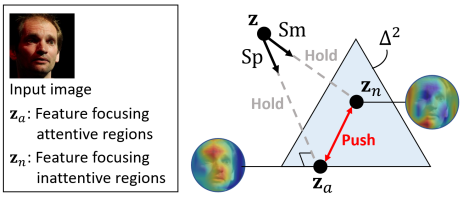

Fig. 4. Conceptual illustration for geometric interpretation of Sp and Sm. The circle-shaped pictures show the face region that each feature pays attention to. Red arrow and gray-dotted line indicate difference in semantic and content information, respectively

where $\mathbf{1}_{pol}$ is a kind of penalty function that has 0 if the valence signs of the two inputs are the same, and -0.5 otherwise. As a result, Eq. (5), which independently describes angle, intensity, and polarity, can cover the universal meaning of emotions [17]. Default setting of the similarity function f is Eq. (4), and the comparison analysis of f_1 and f_2 is dealt with in Section 4.5.

3.5 Feature Transformations for Self-supervision

We adopt SparseMax (Sp) and SoftMax (Sm) [28] to generate attentive and inattentive regions from $\mathbf{z}$ on simplex Δ^{d-1}. Sp, which is in charge of a sparse neural attention mechanism, may be responsible for facial regions related to emotional expression. Since Sm is suitable for weighted aggregation of features [59] and is denser than Sp on Δ^{d-1} [33], it is used as a tool to obtain average attention information of the face. Unlike feature attention modules, Sp and Sm can explicitly get features that are relevant or less relevant to facial expressions. Sp and Sm are defined as follows.

$$\mathrm{Sp}(\mathbf{z}) := \underset{\mathbf{p}\in\Delta^{d-1}}{\arg\max}\, \langle \mathbf{z}, \mathbf{p}\rangle - \frac{1}{2}\|\mathbf{p}\|^2 = \underset{\mathbf{p}\in\Delta^{d-1}}{\arg\min} \|\mathbf{p}-\mathbf{z}\|^2 \tag{6}$$

$$\mathrm{Sm}(\mathbf{z}) := \underset{\mathbf{p}\in\Delta^{d-1}}{\arg\max}\, \langle \mathbf{z}, \mathbf{p}\rangle + \mathcal{H}(\mathbf{p}) = \frac{e^{\mathbf{z}}}{\sum_i e^{\mathbf{z}_i}} \tag{7}$$

where $\Delta^{d-1} = \left\{\mathbf{p}\in\mathbb{R}^d_+ \mid \|\mathbf{p}\|_1 = 1\right\}$, $\mathcal{H}(\mathbf{p}) = -\sum_i p_i \ln p_i$, i.e., the negative Shannon entropy. Eqs. (6) and (7) that return $\mathbf{z}_a$ and $\mathbf{z}_n$ respectively are continuous and differentiable. Since the output of Sp corresponds to Euclidean projection onto the simplex, it is sparse. On the other hand, since $\exp(\cdot) > 0$, the output of Sm is dense. It is noteworthy that $\mathbf{z}_a$ and $\mathbf{z}_n$ are geometrically located on the edge and inside of Δ^2, respectively as in Fig. 4. Δ^2, where a vector in which emotions are strongly expressed is located outside, has a similar structural characteristics to the AV space. Even if dimension d increases ($d > 3$), the high-dimensional $\mathbf{z}_a$ and $\mathbf{z}_n$ on Δ^{d-1} can be projected while maintaining the relationship between Sp and Sm. So, Δ^{d-1} can be considered as the high-dimensional *emotional space*. Note that the projection head (H) for feature transformations plays a role of reducing the dimension from Δ^{d-1} to AV space while maintaining the emotional characteristics between $\mathbf{z}_a$ and $\mathbf{z}_n$.

Implementation. We generate $\mathbf{z}_a$ and $\mathbf{z}_n$ in each forward pass through an iterative optimization process based on the CVXPY library [10]. Specifically, ECOS (embedded conic solver) takes about 1.3 seconds per mini-batch on Xeon(R) E5-1650 CPU to generate $\mathbf{z}_a$ and $\mathbf{z}_n$ (cf. Appendix for forward/backward passes).

3.6 Discriminative Learning

In order to effectively realize the push and pull strategy of CRL, the emotion representation levels of two self-supervisions $\mathbf{z}_a$ and $\mathbf{z}_n$ should be differentiated from each other. At the same time, self-supervisions should preserve the content information of $\mathbf{z}$ to some extent. So, we construct the triplet tuple $Z = (\mathbf{z}, \mathbf{z}_a, \mathbf{z}_n)$, and define a discriminative objective function based on the triplet loss [43] as follows:

$$\mathcal{L}_{dis}(Z) = \sum_{(\mathbf{v},\mathbf{v}_a,\mathbf{v}_n)\in\mathcal{V}} \underbrace{\left[\delta_1 - \|\mathbf{v}_a - \mathbf{v}_n\|^2\right]_+}_{\text{Push term}} + \underbrace{(\|\bar{\mathbf{v}}_a\| - \delta_2) + (\|\bar{\mathbf{v}}_n\| - \delta_2)}_{\text{Hold terms}} \quad (8)$$

where δ_1 indicates the margin of $[\cdot]_+$, and δ_2 is a holding factor. Another projection head (H_1) projects $\mathbf{z}$ into $\mathbf{v}$ on metric space $\mathcal{V}$: $\mathbf{v} = \mathrm{H}_1(\mathbf{z})$. $\bar{\mathbf{v}}_* = \frac{\mathbf{v}_*}{\|\mathbf{v}_*\|} - \frac{\mathbf{v}}{\|\mathbf{v}\|}$ indicates the (unit) vector difference. The first term of Eq. (8) is designed in such a way that $\mathbf{v}_a$ and $\mathbf{v}_n$ push each other within δ_1 so that $\mathcal{L}_{AVCE}$ can learn semantically discriminated views. The remaining terms properly hold the difference between $\mathbf{v}$ and $\mathbf{v}_a$ (or $\mathbf{v}_n$) so that the content information is preserved as shown in Fig. 4. Here, these terms were designed with inspiration from a previous study [51] where a powerful transformation (or augmentation) would be possible if content preserving semantic transformations were allowed.

Algorithm 1 Training Procedure of AVCE

Require: Input image IMG, learning rate ϵ_1, ϵ_2, ground-truth GT, parameters of E, C, R, H, H_1 (θ_e, θ_c, θ_r, θ_h, θ_{h1}).
Ensure: Initialize (θ_e, θ_c, θ_r, θ_h, θ_{h1}) to Normal distribution.
  while not converge (θ_e, θ_c, θ_r, θ_h, θ_{h1}) **do**
    (Forward pass 1)
    $\mathbf{z} = (\mathrm{C} \circ \mathrm{E})(\mathrm{IMG})$ ▷ Input image encoding
    $\mathbf{x} = \mathrm{R}(\mathbf{z})$
    $\mathbf{z}_a, \mathbf{z}_n = \mathrm{Sp}(\mathbf{z}), \mathrm{Sm}(\mathbf{z})$ ▷ Feature transformations
    $\mathbf{y}_a, \mathbf{y}_n = \mathrm{H}(\mathbf{z}_a), \mathrm{H}(\mathbf{z}_n)$
    $\mathcal{L}_{main} \leftarrow \mathcal{L}_{AV}(\mathbf{x},\mathrm{GT}) + \mathcal{L}_{AVCE}(\mathbf{x},\mathbf{y}_a,\mathbf{y}_n)$
    (Backward pass 1)
    $(\theta_e, \theta_c, \theta_r, \theta_h) \leftarrow (\theta_e, \theta_c, \theta_r, \theta_h) - \epsilon_1 \nabla_{(\theta_e,\theta_c,\theta_r,\theta_h)} \mathcal{L}_{main}$
    (Forward pass 2)
    $\mathcal{L}_{dis} \leftarrow \mathrm{Triplet}(\mathrm{H}_1(\mathbf{z}, \mathbf{z}_a, \mathbf{z}_n))$ ▷ Eq. (8)
    (Backward pass 2)
    $(\theta_e, \theta_c, \theta_{h1}) \leftarrow (\theta_e, \theta_c, \theta_{h1}) - \epsilon_2 \nabla_{(\theta_e,\theta_c,\theta_{h1})} \mathcal{L}_{dis}$
  end while

Training procedure. The proposed method performs *two* forward/backward passes every iteration. Algorithm 1 describes the details of the objective functions calculated at each step and the neural networks to be trained. In the first pass, the encoding process of an input image and the feature transformation process are performed. Then, GT-based supervised learning ($\mathcal{L}_{AV}$) and contrastive learning ($\mathcal{L}_{AVCE}$) based on self-supervisions $\mathbf{y}_a$ and $\mathbf{y}_n$ are performed, respectively. These two loss functions are merged into $\mathcal{L}_{main}$ to update the trainable parameters ($\theta_e, \theta_c, \theta_r, \theta_h$). In the second pass, discriminative learning ($\mathcal{L}_{dis}$) with triplet tuple Z as input is performed and parameters ($\theta_e, \theta_c, \theta_{h1}$) are updated.

4 Experiments

4.1 Datasets

We adopted open datasets only for research purposes, and informed consent was obtained if necessary. AFEW-VA [24] derived from the AFEW dataset [9] consists of about 600 short video clips annotated with AV labels frame by frame. Like [19], evaluation was performed through cross validation at a ratio of 5:1. Aff-wild [55] consists of about 300 video clips obtained from various subjects watching movies and TV shows. Since the test data of the Aff-wild was not disclosed, this paper adopted the sampled train set for evaluation purpose in the same way as previous works [14,19]. Aff-wild2 [22] is a dataset in which about 80 training videos and about 70 evaluation videos are added to Aff-wild to account for spontaneous facial expressions. AffectNet [31] consists of about 440K static images annotated with AV and discrete emotion labels, and landmarks.

4.2 Configurations

All networks were implemented in PyTorch, and the following experiments were performed on Intel Xeon CPU and RTX 3090 GPU. Each experiment was repeated five times. Encoder (E) was designed with parameter-reduced AlexNet [25] and ResNet18 [15] from scratch. Compressor (C) is composed of average pooling and FC layer. H($/\mathrm{H}_1$) and R are composed of FC layers and batch normalization (cf. Appendix for the network details). Adam optimizer [20] with a learning rate (LR) of 1e-4 was used to optimize E, C, and R. SGD [39] with LR 1e-2 was used to optimize H and H_1. AFEW-VA and AffectNet were trained for 50K iterations, and Aff-wild(/2) was trained for 100K iterations. Here, LR was reduced by 0.8 times at the initial 5K iterations, and decreased by 0.8 times every 20K iterations. The mini-batch sizes of AlexNet (AL) and ResNet18 (R18) were set to 256 and 128, respectively.

Hyperparameters. For face detection, the latest version of deep face detector [57] was used, and the detected facial regions were resized to 224×224 through random cropping (center cropping when testing). The dimensions of $\mathbf{z}$ and $\mathbf{v}$ were set to 32 and 8, respectively. In Eq. (2), α, β, and γ were set to 0.5, 0.005, and 0.5 according to Lemma 1. Angular similarity of Eq. (2) was clipped to lie

in the range $(0, 1]$. In Eq. (4), the balance factor μ was set to 0.75. In Eq. (8), the margin δ_1 of the hinge function $[\cdot]_+$ and holding factor δ_2 were set to 1.0 and 0.8, respectively.

Evaluation metrics. Root mean-squared error (RMSE) and sign agreement (SAGR) were used to measure the point-wise difference and overall emotional degree. In addition, Pearson correlation coefficient (PCC) and concordance correlation coefficient (CCC) were employed to measure the emotion tendency. For details on the above metrics, refer to the Appendix. Since the objective of emotional learning is to simultaneously achieve the minimization of RMSE and the maximization of PCC/CCC [19,23], $\mathcal{L}_{AV}$ is designed as follows: $\mathcal{L}_{AV} = \mathcal{L}_{RMSE} + \frac{(\mathcal{L}_{PCC}+\mathcal{L}_{CCC})}{2}$. Here, $\mathcal{L}_{C(/P)CC} = 1 - \frac{C(/P)CC_a + C(/P)CC_v}{2}$.

Table 1. Comparison results on the AFEW-VA dataset. Red and blue indicate the first and second-ranked values, respectively. For all comparison methods, the numerical values specified in the paper were used as they are

Case	Methods	RMSE (↓)		SAGR (↑)		PCC (↑)		CCC (↑)	
		(V)	(A)	(V)	(A)	(V)	(A)	(V)	(A)
Static	Kossaifi et al. [24]	0.27	0.23	-	-	0.41	0.45	-	-
	Mitenkova et al. [30]	0.40	0.41	-	-	0.33	0.42	0.33	0.40
	Kossaifi et al. [23]	0.24	0.24	0.64	0.77	0.55	0.57	0.55	0.52
	CAF (R18) [19]	0.17	0.18	0.68	0.87	0.67	0.60	0.59	0.54
	CAF (AL) [19]	0.20	0.20	0.66	0.83	0.67	0.63	0.58	0.57
	AVCE (R18) (Ours)	0.156	0.144	0.783	0.876	0.651	0.727	0.619	0.707
	AVCE (AL) (Ours)	0.162	0.170	0.790	0.834	0.730	0.686	0.629	0.622
Temporal	Kollias et al. [21]	-	-	-	-	0.51	0.58	0.52	0.56
	Kossaifi et al. [23]-scratch	0.28	0.19	0.53	0.75	0.12	0.23	0.11	0.15
	Kossaifi et al. [23]-trans.	0.20	0.21	0.67	0.79	0.64	0.62	0.57	0.56

4.3 Quantitative Analysis

This section demonstrated the superiority of AVCE by comparing with the latest AV-based FER methods [19,14,23,42] which were verified in the wild datasets. Table 1 showed that AVCE outperforms other methods for AFEW-VA. This is because AVCE can discern even the subtle differences between positive and negative emotions. For example, AVCE (AL) showed about 0.13 higher SAGR (V) and about 0.05 higher CCC (V) than CAF (AL) [19], i.e., the runner-up method.

Next, AVCE showed a noticeable improvement in terms of PCC/CCC compared to CAF for Aff-wild. In Table 2, AVCE (AL) showed about 0.16 (16%) higher PCC (V) and about 0.14 (14%) higher CCC (V) than CAF (AL). Meanwhile, RMSE, which indicates the precision of prediction, was generally superior in R18, and PCC/CCC, which indicates the tendency of emotional change, showed superiority in AL. This tendency demonstrates that CNNs can improve precision in most over-parameterized settings, but CNNs are seldom generalized.

Table 2. Comparison results on the Aff-wild dataset. * was evaluated on Aff-wild's test set using ResNet50 backbone

Methods	RMSE (↓)		SAGR (↑)		PCC (↑)		CCC (↑)	
	(V)	(A)	(V)	(A)	(V)	(A)	(V)	(A)
Hasani et al. [13]	0.27	0.36	0.57	0.74	0.44	0.26	0.36	0.19
Hasani et al. [14]	0.26	0.31	0.77	0.75	0.42	0.40	0.37	0.31
Deng et al. [8]*	-	-	-	-	-	-	0.58	0.52
CAF (R18) [19]	0.22	0.20	0.70	0.76	0.57	0.57	0.55	0.56
CAF (AL) [19]	0.24	0.21	0.68	0.78	0.55	0.57	0.54	0.56
AVCE (R18) (Ours)	0.148	0.152	0.798	0.781	0.600	0.621	0.552	0.583
AVCE (AL) (Ours)	0.154	0.154	0.849	0.795	0.713	0.632	0.682	0.594

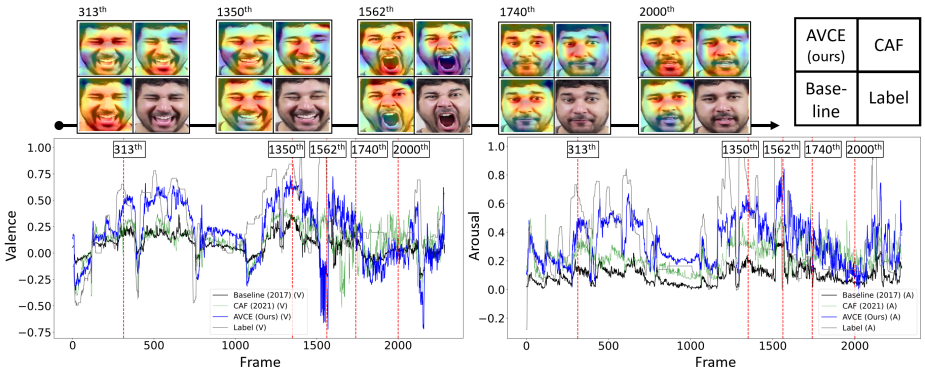

Fig. 5. Analysis of frame unit emotional fluctuations and corresponding mean neural activation maps on Aff-wild dataset. Baseline [31] and CAF [19] are reproduced for a fair experimental setup. Best viewed in color

Note that even in the Aff-wild2 dataset with various backgrounds and subjects added, AVCE (R18) showed 0.031 higher mean CCC than Sanchez et al. [42], that is the latest SOTA (see Table 3). Finally, AVCE shows superiority in both performance and network size on the AffectNet dataset. Please refer to the Appendix for the AffectNet results, additional backbone results, etc.

4.4 Qualitative Analysis

This section visualizes the performance of AVCE through neural activation map [58]. The activation map is computed from the feature maps and the weight matrices of the last layer of R(or H). Since it is important to consider both arousal and valence to capture emotional attention [18], we observed facial regions associated with emotional expression by averaging the two maps. Various examples of each of A and V are provided in the Appendix.

Figure 5 analyzes frame-by-frame emotional fluctuation by adopting CAF and baseline [31]. Overall, AVCE can successfully capture not only positive peaks

Fig. 6. Influence analysis of self-supervision through mean neural activation map on AffectNet. Best viewed in color

but also negative changes. Seeing the 1562-th frame of the left (valence), the activation map of AVCE correctly captured the eye and lip regions and showed significant valence fluctuations. However, this variation showed the opposite direction to the GT. This indicates that it is sometimes difficult to grasp the global semantic context of the video clip only with a single frame.

In addition, we compared the activation maps of R and H to indirectly verify the effect of Sp, which is difficult to visualize. In Fig. 6, AVCE (H), which is trained to encourage the function of Sp, captured emotion-related regions well showing sparser results than AVCE (R). Through the examples in Fig. 6, we can find that the proposed method captures core regions (e.g. eyes and mouth) for FER better than other methods.

Table 3. Results on the validation set of Aff-wild2

Methods	CCC (V)	CCC (A)	Mean
ConvGRU [5]	0.398	0.503	0.450
Self-Attention [50]	0.419	0.505	0.462
Sanchez et al. [42]	0.438	0.498	0.468
AVCE (R18)	0.484	0.513	0.499
AVCE (AL)	0.496	0.500	0.498

Table 4. Ablation study on Aff-wild

CRL formula	f	f_1	f_2	CCC (↑) (V)	CCC (↑) (A)
InfoNCE [35]	✓			0.637	0.546
			✓	0.651	0.550
Barlow-Twins [56]	✓			0.653	0.566
			✓	0.656	0.554
AVCE (AL) w/o $\mathcal{L}_{dis}$	✓			0.642	0.548
AVCE (AL)	✓			0.682	0.594
		✓		0.640	0.577
			✓	0.691	0.581

4.5 Ablation Study

Table 4 further analyzed the superiority of AVCE through representative CRL formulas and similarity functions of Section 3.4. InfoNCE (Eq. (1)) showed worse CCC (V) by 0.045 than AVCE (AL). This gap is lower than when $\mathcal{L}_{dis}$ was not used. Even a cutting-edge Barlow-Twins [56] showed 0.029 worse CCC (V) than AVCE. This proves the strength of AVCE, which reflects the structural property of the AV space well. On the other hand, f_1 based on dot product showed 0.042 lower CCC (V) than f. f_2, which gives a penalty on the valence axis, showed 0.009 high CCC (V) in AVCE, but decreased 0.013 in the arousal

axis. In addition, the mining method [6] used as post-processing for negative pair sampling of AVCE shows an improvement of about 0.02 in terms of CCC (A). For details of the inference speed and the impact of AVCE, refer to the Appendix.

Voting results of user study. Finally, we conducted a user study to validate the emotion-aware ability of feature transformations. For this experiment, we prepared 32 pairs of examples generated by neural activation maps based on the same input image (cf. Figs. 3 and 4 in Appendix). For each example, 12 subjects were instructed to rank the images in the order of the best captures of emotional expression. As a result, AVCE, CAF, and baseline [31] showed top-1 accuracy of 67.96%, 25.78%, and 6.26%, respectively. Therefore, the superiority of the proposed method was proven through this user study once again.

5 Discussion of Limitations

Network design. One may argue about the use of spatio-temporal network such as [42]. However, all methods showing excellent performance are based on static images (cf. Table 1). This shows that emotional expression-aware self-supervision, that is, attentive region of AVCE, is a more important clue for AV-based FER than quantitative differences in temporal features so far.

Data imbalance. Since the datasets used in the experiments are biased towards positive emotions, training neural networks with only GT causes a bias towards positive emotions. In the future, study on weighted resampling or distribution shift that can explicitly deal with this data imbalance issue should be done.

Other risk factors. AV-based FER should be robust against both internal factors (e.g. skin color, face angle) and external factors (e.g. illumination and background) of a subject. This paper used datasets containing internal factors of various properties for learning, but did not directly focus on external factors. In the future, illumination and backgrounds-aware attention ability should be additionally considered in AVCE.

6 Conclusion

For the first time in the AV-based FER field, we presented a self-supervised method to learn emotion-aware facial features. Thanks to the features obtained from the novel iterative process, the proposed AVCE can understand the emotions from various perspectives. Experiments show that AVCE can detect core regions of wild facial emotions and regress continuous emotional changes without temporal learning.

Acknowledgements This work was supported by IITP grants funded by the Korea government (MSIT) (No. 2021-0-02068, AI Innovation Hub and RS-2022-00155915, Artificial Intelligence Convergence Research Center(Inha University)), and was supported by the NRF grant funded by the Korea government (MSIT) (No. 2022R1A2C2010095 and No. 2022R1A4A1033549).

References

1. Barros, P., Parisi, G., Wermter, S.: A personalized affective memory model for improving emotion recognition. In: International Conference on Machine Learning. pp. 485–494 (2019)
2. Bera, A., Randhavane, T., Manocha, D.: Modelling multi-channel emotions using facial expression and trajectory cues for improving socially-aware robot navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
3. Cerf, M., Frady, E.P., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of vision **9**(12), 10–10 (2009)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607 (2020)
5. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, Doha, Qatar. pp. 1724–1734. ACL (2014)
6. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 8765–8775. Curran Associates, Inc. (2020)
7. d'Apolito, S., Paudel, D.P., Huang, Z., Romero, A., Van Gool, L.: Ganmut: Learning interpretable conditional space for gamut of emotions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 568–577 (2021)
8. Deng, D., Chen, Z., Zhou, Y., Shi, B.: Mimamo net: Integrating micro-and macro-motion for video emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 2621–2628 (2020)
9. Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 653–656 (2018)
10. Diamond, S., Boyd, S.: Cvxpy: A python-embedded modeling language for convex optimization. The Journal of Machine Learning Research **17**(1), 2909–2913 (2016)
11. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018)
12. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
13. Hasani, B., Mahoor, M.H.: Facial affect estimation in the wild using deep residual and convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 9–16 (2017)
14. Hasani, B., Negi, P.S., Mahoor, M.: Breg-next: Facial affect computing using adaptive residual networks with bounded gradient. IEEE Transactions on Affective Computing (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

16. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews neuroscience **2**(3), 194–203 (2001)
17. Jackson, J.C., Watts, J., Henry, T.R., List, J.M., Forkel, R., Mucha, P.J., Greenhill, S.J., Gray, R.D., Lindquist, K.A.: Emotion semantics show both cultural variation and universal structure. Science **366**(6472), 1517–1522 (2019)
18. Jefferies, L.N., Smilek, D., Eich, E., Enns, J.T.: Emotional valence and arousal interact in attentional control. Psychological science **19**(3), 290–295 (2008)
19. Kim, D.H., Song, B.C.: Contrastive adversarial learning for person independent facial emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5948–5956 (2021)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
21. Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. International Journal of Computer Vision **127**(6-7), 907–929 (2019)
22. Kollias, D., Zafeiriou, S.: Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. In: 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019. p. 297 (2019), https://bmvc2019.org/wp-content/uploads/papers/0399-paper.pdf
23. Kossaifi, J., Toisoul, A., Bulat, A., Panagakis, Y., Hospedales, T.M., Pantic, M.: Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6060–6069 (2020)
24. Kossaifi, J., Tzimiropoulos, G., Todorovic, S., Pantic, M.: Afew-va database for valence and arousal estimation in-the-wild. Image and Vision Computing **65**, 23–36 (2017)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
26. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering (2021)
27. Marinoiu, E., Zanfir, M., Olaru, V., Sminchisescu, C.: 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2158–2167 (2018)
28. Martins, A., Astudillo, R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In: International conference on machine learning. pp. 1614–1623 (2016)
29. Mikels, J.A., Fredrickson, B.L., Larkin, G.R., Lindberg, C.M., Maglio, S.J., Reuter-Lorenz, P.A.: Emotional category data on images from the international affective picture system. Behavior research methods **37**(4), 626–630 (2005)
30. Mitenkova, A., Kossaifi, J., Panagakis, Y., Pantic, M.: Valence and arousal estimation in-the-wild with tensor methods. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–7 (2019)
31. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)

32. Mroueh, Y., Melnyk, I., Dognin, P., Ross, J., Sercu, T.: Improved mutual information estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9009–9017 (2021)
33. Niculae, V., Martins, A., Blondel, M., Cardie, C.: Sparsemap: Differentiable sparse structured inference. In: International Conference on Machine Learning. pp. 3799–3808 (2018)
34. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016)
35. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
36. Panda, R., Zhang, J., Li, H., Lee, J.Y., Lu, X., Roy-Chowdhury, A.K.: Contemplating visual emotions: Understanding and overcoming dataset bias. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 579–595 (2018)
37. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology **17**(3), 715–734 (2005)
38. Rafaeli, A., Sutton, R.I.: Emotional contrast strategies as means of social influence: Lessons from criminal interrogators and bill collectors. Academy of management journal **34**(4), 749–775 (1991)
39. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics pp. 400–407 (1951)
40. Roy, S., Etemad, A.: Self-supervised contrastive learning of multi-view facial expressions. arXiv preprint arXiv:2108.06723 (2021)
41. Roy, S., Etemad, A.: Spatiotemporal contrastive learning of facial expressions in videos. arXiv preprint arXiv:2108.03064 (2021)
42. Sanchez, E., Tellamekala, M.K., Valstar, M., Tzimiropoulos, G.: Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9074–9084 (2021)
43. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
44. Song, B.C., Kim, D.H.: Hidden emotion detection using multi-modal signals. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–7 (2021)
45. Srivastava, P., Srinivasan, N.: Time course of visual attention with emotional faces. Attention, Perception, & Psychophysics **72**(2), 369–377 (2010)
46. Taverner, J., Vivancos, E., Botti, V.: A multidimensional culturally adapted representation of emotions for affective computational simulation and recognition. IEEE Transactions on Affective Computing (2020)
47. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 776–794. Springer (2020)
48. Tsai, Y.H., Zhao, H., Yamada, M., Morency, L.P., Salakhutdinov, R.: Neural methods for point-wise dependency estimation. In: Proceedings of the Neural Information Processing Systems Conference (Neurips) (2020)
49. Tsai, Y.H.H., Ma, M.Q., Yang, M., Zhao, H., Morency, L.P., Salakhutdinov, R.: Self-supervised representation learning with relative predictive coding. In: International Conference on Learning Representations (2021)

50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30, pp. 5998–6008. Curran Associates, Inc. (2017)
51. Wang, Y., Pan, X., Song, S., Zhang, H., Huang, G., Wu, C.: Implicit semantic data augmentation for deep networks. Advances in Neural Information Processing Systems **32**, 12635–12644 (2019)
52. Wei, Z., Zhang, J., Lin, Z., Lee, J.Y., Balasubramanian, N., Hoai, M., Samaras, D.: Learning visual emotion representations from web data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13106–13115 (2020)
53. Xue, F., Wang, Q., Guo, G.: Transfer: Learning relation-aware facial expression representations with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3601–3610 (2021)
54. Yang, J., Li, J., Li, L., Wang, X., Gao, X.: A circular-structured representation for visual emotion distribution learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4237–4246 (2021)
55. Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-wild: Valence and arousal'in-the-wild'challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 34–41 (2017)
56. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Proceedings of the 38th International Conference on Machine Learning, Virtual Event. vol. 139, pp. 12310–12320. PMLR (2021)
57. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
58. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
59. Zhu, X., Xu, C., Tao, D.: Where and what? examining interpretable disentangled representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5861–5870 (2021)