Contrastive Adversarial Learning for Person Independent Facial Emotion Recognition

Daeha Kim, Byung Cheol Song
Department of Electronic Engineering, Inha University, Incheon 22212, South Korea
kdhht5022@gmail.com, bcsong@inha.ac.kr

Abstract

Since most facial emotion recognition (FER) methods significantly rely on supervision information, they have a limit to analyzing emotions independently of persons. On the other hand, adversarial learning is a well-known approach for generalized representation learning because it never requires supervision information. This paper presents a new adversarial learning for FER. In detail, the proposed learning enables the FER network to better understand complex emotional elements inherent in strong emotions by adversarially learning weak emotion samples based on strong emotion samples. As a result, the proposed method can recognize the emotions independently of persons because it understands facial expressions more accurately. In addition, we propose a contrastive loss function for efficient adversarial learning. Finally, the proposed adversarial learning scheme was theoretically verified, and it was experimentally proven to show state of the art (SOTA) performance.

Introduction

With rapid development of deep learning, facial emotion recognition (FER) is being studied more and more actively. Until now, FER based on several discrete emotion categories has been dominantly developed (Vielzeuf, Pateux, and Jurie 2017). However, discrete domain FER is limited in capturing the rich emotions of a person. For example, although previous approach (Harmon-Jones et al. 2011) could judge that a certain person's emotion was anger, it could not catch the strength of the emotion.

Arousal-valence (AV) domain FER technology which represents and analyzes more sophisticated emotions are receiving much attention (Zafeiriou et al. 2017). Arousal indicates how active the emotion is, and valence indicates how positive or negative the emotion is. Note that AV labels are classified as a regression task because they have continuous values. Recently, several continuous domain datasets have been released, and AV domain FER methods for the datasets have been proposed (Mollahosseini, Hasani, and Mahoor 2017; Kossaifi et al. 2017).

However, conventional FER methods in AV domain have a critical disadvantage. Since they (Kossaifi et al. 2020; Hasani, Negi, and Mahoor 2020) depend on supervision information, they tend to be biased toward given data, i.e., person-dependent. In particular, this phenomenon often occurs in complex or strong facial expressions (Zeng, Shan, and Chen 2018). For example, when a person's emotion is angry, meaningful changes occur in the lips and eyes, etc. This change may vary depending on the intrinsic mood or the other person's emotions (Russell 2017). So, successfully capturing the diversity of such complex emotions is the key to the person-independent FER. However, if the compressed characteristics of persons/subjects is simply utilized through a deep network, diversity analysis will become difficult. Therefore, an advanced learning for dealing with diversity analysis plays an important role in designing a person-independent FER model.

Inspired from (Goodfellow et al. 2014), this paper proposes a new structure that learns the difference in emotional intensity adversarially, and adopts the learned information as auxiliary data for FER in the AV domain. Since the main goal of adversarial learning is to understand the characteristics of a target distribution, adversarial learning enables the proposed network to better understand the embedding representation of strong or complex emotions (Yang, Ciftci, and Yin 2018). Thus, the proposed method is not only based on the AV domain but also is independent of persons, so it can recognize more sophisticated and generalized emotions.

Specifically, the proposed method consists of three steps as shown in Figure 1: 1) Emotional binarization process based on Otsu thresholding (arrow ①), 2) adversarial learning based on critic network in latent feature space (arrows ② and ③), 3) AV domain emotion learning (arrow ④). Here, the strong and weak emotion image sets generated through the binarization process include both facial expressions as well as inner emotions, which stands out in relatively strong emotions (Balconi, Vanutelli, and Finocchiaro 2014). In other words, changes in inner emotions that are not revealed in facial expressions, such as hidden emotions, can also be used for adversarial learning.

On the other hand, the loss function of the critic network can be defined using Maximum Mean Discrepancy (MMD) (Gretton et al. 2012) or Integral Probability Metric (IPM) (Mroueh, Sercu, and Goel 2017), which quantifies and dis-

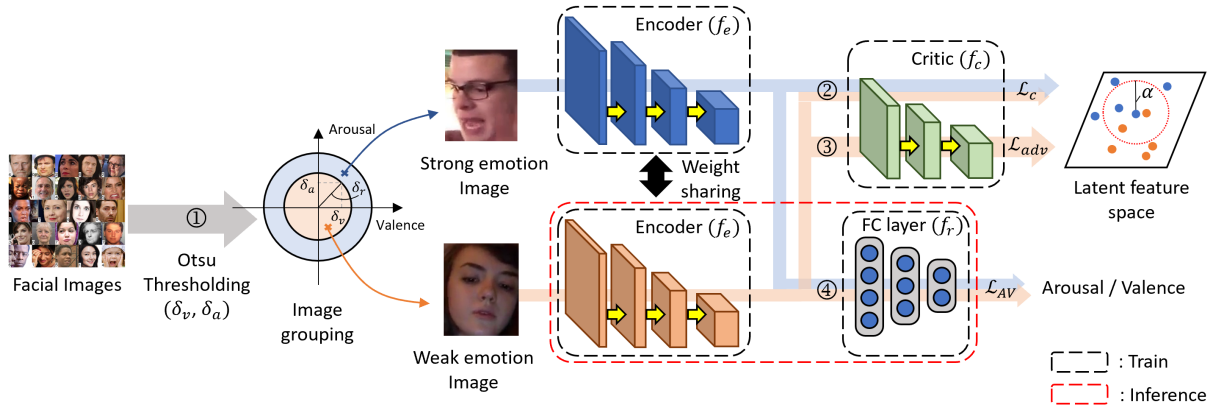

Figure 1: The overview of the proposed method. ② means discriminative learning of the loss function $\mathcal{L}_c$ defined by the distributions of the two groups and margin α, and ③ means adversarial learning of the loss function $\mathcal{L}_{adv}$ defined only by the weak emotional distribution. Here, the learnable parameters of the critic network are frozen. ④ indicates AV domain regression of the loss function $\mathcal{L}_{AV}$ defined by the output of FC layer.

criminates the difference in distribution. However, since these metrics simply learn the two distributions in terms of the difference of statistical means, they have limitations in quantifying the embedding representation of complex emotion elements (Arjovsky, Chintala, and Bottou 2017). Thus, we define the loss function of the critic network by using the contrastive loss function (Hadsell, Chopra, and LeCun 2006), which is useful for learning the distribution or relationship between individual samples and variational divergence minimization (VDM) (Nowozin, Cseke, and Tomioka 2016). Like IPM, VDM can quantify the distribution difference and is considered a general form of adversarial learning (Nowozin, Cseke, and Tomioka 2016). Also, since the contrastive loss function is computed using samples within a margin, it can learn the features of the local distribution.

The contribution points of this paper are as follows.

- A new adversarial learning structure based on theoretical analysis presents a new research orientation of continuous domain FER.
- Although the proposed method does not utilize temporal information, it shows significantly better performance than conventional schemes in video as well as static datasets.

Related Work

Facial emotion recognition in AV domain: The basic approach to AV domain FER (Mollahosseini, Hasani, and Mahoor 2017) is to solve a regression problem based on convolutional neural networks (CNNs) such as AlexNet (Krizhevsky, Sutskever, and Hinton 2012) and ResNet (He et al. 2016). It was reported that multiple tasks of face detection, smile prediction, and AV estimation could be simultaneously performed via fusion of a detection network and a regression network (Jang, Gunes, and Patras 2019). Recently, Kossaifi et al. proposed a factorized higher-order CNN (FHC) that efficiently learns intra-frame and inter-frame differences through decomposition of a convolution operation (Kossaifi et al. 2020). However, since FHC focused on CNN computation rather than FER, it did not show state-of-the-art (SOTA) performance. The latest BreG-NeXt (Hasani, Negi, and Mahoor 2020) improved the residual mapping structure so that it is not only light in terms of computational cost, but also shows SOTA performance. BreG-NeXt is getting a lot of attention because it performs best in both the discrete and continuous domain FER.

Since conventional methods including BreG-NeXt were designed based on direct matching between input images and supervision information, they could learn even non-facial features. In other words, they can depend on persons rather than facial expressions by themselves. Some studies tried to recognize emotions independently of persons by employing an adversarial learning structure (Zhang et al. 2018; Cai et al. 2019). But, they were targeting at discrete FER.

Adversarial learning: Adversarial learning (Goodfellow et al. 2014) emphasizes the roles of generation and discrimination tasks through adversarial relationship between generator and critic network. However, adversarial learning may suffer from instability such as mode collapse, and is not scale-invariant. As an approach to solve this problem, various methods to effectively train the critic network have been developed (Srivastava et al. 2019; Mroueh and Sercu 2017; Nowozin, Cseke, and Tomioka 2016). FisherGAN (Mroueh and Sercu 2017), which was inspired by Fisher discriminant analysis and showed strength in terms of learning stability and efficient computation, is a representative case where IPM is used as a loss function of the critic network. GramNet (Srivastava et al. 2019) adopted the difference in squared ratio as a loss function based on the characteristics of MMD. Using a variational method (Nguyen, Wainwright, and Jordan 2010) to estimate φ-divergence, a VDM structure for adversarial learning was proposed in (Nowozin, Cseke, and Tomioka 2016).

Adversarial learning is less dependent on data because it learns the relationship of latent features rather than supervision information. Thus, it can even analyze complex emo-

tions that supervision information cannot represent.

Deep metric learning: Deep metric learning (DML) is a method of learning the sample similarity through a specific metric such as Euclidean distance. For instance, contrastive loss and triplet loss are to learn similarity based on a pair of samples (Schroff, Kalenichenko, and Philbin 2015). If a pair has different class labels, the metric loss is learned so that they are separated from each other, and vice versa.

Method

Notation and Fisher IPM

This section describes a principle of discriminative learning in the critic network through Fisher IPM (Mroueh and Sercu 2017). Assume that a set of n-dimensional facial images and their AV labels are defined as $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$. The cardinality of this set is N. Each AV label consists of arousal y_a and valence y_v. Let $\mathbf{z}_i = (f_c \circ f_e)(\mathbf{x}_i) \subset \mathcal{Z}$ denote the d-dimensional latent feature obtained from the i-th facial image $\mathbf{x}_i$. Here, f_e and f_c denote the encoder and the critic network, respectively. Also, let $\mathcal{F}$ be a set of measurable and bounded real-valued functions on compact space $\mathcal{Z}$. If $\mathbb{P}$ and $\mathbb{Q}$ are given through two sets of arbitrary random probability distributions on $\mathcal{Z}$, the Fisher IPM by $\mathcal{F}$ is defined by

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \{\mathbb{E}_{\mathbf{z}\sim\mathbb{P}} f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}} f(\mathbf{z})\} \tag{1}$$

For convenience, the 2nd order regularization term of f is omitted. Assuming that $\mathbb{P}$ and $\mathbb{Q}$ are distributions of strong and weak emotions, respectively, the critic network can be learned to discriminate the emotional intensity through emotion vectors $\mathbf{z}_s$ and $\mathbf{z}_w$ obtained from $\mathbb{P}$ and $\mathbb{Q}$.

Contrastive Adversarial Framework

The goal of the proposed method is to accomplish person-independent learning by using discriminative learning of the critic network and adversarial learning of the encoder network as auxiliary information. First, design a loss function for discriminative learning through φ-divergence that quantifies distribution differences. The lower bound induced by φ-divergence and variational method is defined as follows (Nguyen, Wainwright, and Jordan 2010):

$$\begin{aligned} d_{\varphi}(\mathbb{P}, \mathbb{Q}) &= \int_{\mathcal{Z}} \varphi\left(\frac{\mathbb{P}(\mathbf{z})}{\mathbb{Q}(\mathbf{z})}\right) \mathbb{Q}(\mathbf{z})\, d\mathbf{z} \\ &\geq \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}} f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}} \varphi^*(f(\mathbf{z})) \end{aligned} \tag{2}$$

where $\varphi : \mathbb{R}^+ \to \mathbb{R}$ is a convex, lower semi-continuous function satisfying $\varphi(1) = 0$ and φ^* indicates Fenchel conjugate of φ (Hiriart-Urruty and Lemaréchal 2012). The lower bound in Eq. (2) is equivalent to the shape of VDM in (Nowozin, Cseke, and Tomioka 2016). Similar to Fisher IPM, the lower bound can be learned to discriminate two different distributions. If appropriate φ and upper bound d_{φ} are chosen, the loss function can be computed within the bounds guaranteed. Next, determine an appropriate upper bound via Lemma 1 and transform the lower bound of d_{φ}.

Lemma 1. When d_{φ} in Eq. (2) is Chi-squared (χ_2) distance and the corresponding φ is properly chosen, Eq. (2) is redefined as follows:

$$\begin{aligned} &d_{\varphi}(\mathbb{P}, \mathbb{Q}) \\ &\geq \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}} f(\mathbf{z}) - \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim(\mathbb{P}+\mathbb{Q})} f(\mathbf{z}) - \frac{1}{4}\mathbb{E}_{\mathbf{z}\sim\frac{\mathbb{P}+\mathbb{Q}}{2}} f^2(\mathbf{z}) \end{aligned} \tag{3}$$

Proof. Please refer to section C of the supplementary material of (Mroueh and Sercu 2017).

The first term of RHS of Eq. (3) consists of vectors obtained from $\mathbb{P}$, and the second term consists of vectors obtained from $\mathbb{P}$ and $\mathbb{Q}$. The third term can be regarded as a regularization term as in (Mroueh and Sercu 2017). If a function f of $\mathcal{F}$ is a contrastive loss (Hadsell, Chopra, and LeCun 2006) as in Eq. (4), the critic network can be trained so that it increases not only the intra-class compactness of similar vectors from $\mathbb{P}$ but also inter-class variability of different vectors from $\mathbb{P}$ and $\mathbb{Q}$.

$$f_{cont}(\mathbf{z}_i, \mathbf{z}_j) = -\xi_{i,j} \max\left\{0, \xi_{i,j}\left(D^2(\mathbf{z}_i, \mathbf{z}_j) - \alpha\right)\right\} \tag{4}$$

where $\max\{0, \cdot\}$ is the hinge function, D is the Euclidean distance, and α is the margin (as with the contrastive loss, α is applied only to the second term of Eq. (3)). As a group indicator, $\xi_{i,j}$ is set to 1 if $\mathbf{z}_i$ and $\mathbf{z}_j$ belong to the same group, and -1 otherwise.

Therefore, discriminative learning of the critic network is performed with the RHS of Eq. (3) as the loss function $\mathcal{L}_c$ (line 6 of Algorithm 1). Note that the RHS of Eq. (3) satisfies $f = f_{cont}$ and has χ_2 distance. Based on (Nguyen, Wainwright, and Jordan 2005), which analyzed the relationship between φ-divergence and hinge function, it is theoretically reasonable to include f_{cont} of Eq. (4) in $\mathcal{F}$ of Eq. (3).

On the other hand, adversarial learning of the encoder network is based on the average of latent features as in (Mroueh and Sercu 2017; Arjovsky, Chintala, and Bottou 2017): $\mathcal{L}_{adv} = -\frac{1}{N_w}\sum_{i=1}^{N_w} \mathbf{z}_{w_i}$ (line 7 of Algorithm 1). Here, N_w denotes the number of weak emotional samples in the minibatch, and when $\mathcal{L}_{adv}$ is learned, the parameter update of the critic network is temporarily stopped.

The critic network and the encoder network are alternately trained. The encoder network trained with the help of the critic network learns AV domain emotions together with the FC layer (line 8 of Algorithm 1). This adversarial learning structure is called the contrastive adversarial framework (CAF) in this paper.

However, Eq. (4) that cannot reflect the difficulty of learning in the latent space can cause overfitting phenomenon in the later stage of learning. The adaptive margin to solve this problem is designed as follows.

CAF with adaptive margin: The basic idea of designing an adaptive margin is to jointly consider the dependence of two groups in the latent space and the learning difficulty. The dependence of two distributions is determined through mutual information (MI) (Belghazi et al. 2018), and the learning difficulty is determined by the confidence interval Δ (Bal-

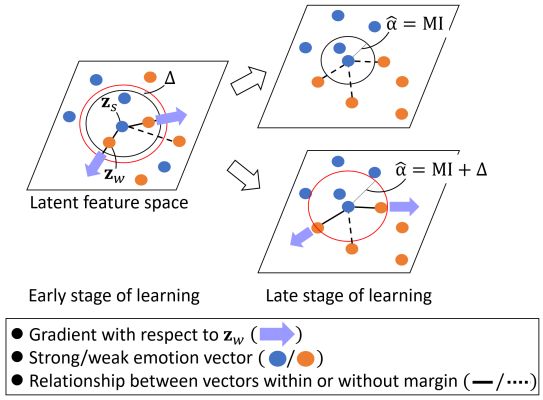

Figure 2: The conceptual analysis of the impact of each factor of adaptive margin on discriminative learning. Black circles mean only MI is considered, and red circles mean MI and confidence intervals are considered together.

subramani et al. 2019). As a result, the adaptive margin is proposed by:

$$\widehat{\alpha} = \mathrm{MI}(g(Z_s), g(Z_w)) + \Delta \tag{5}$$

where Z_s and Z_w denote matrices composed of sets of $\mathbf{z}_s$ and $\mathbf{z}_w$, respectively. Gaussian kernel $g(\cdot)$ is applied so that the joint distribution of Z_s and Z_w becomes a normal distribution.

In Eq. (5), MI provides an appropriate difficulty level for learning the critic network by reflecting the dependence of Z_s and Z_w. Δ defines the difficulty in the latent space as empirical bias, which increases as the proportion of samples with the same class label in the margin area increases (Balsubramani et al. 2019). That is, it gradually increases as learning progresses. Therefore, Δ provides an additional lower bound to $\widehat{\alpha}$ when MI is intractable in the later stage of learning, especially when MI is close to 0.

Figure 2 conceptually describes the effect of each element of $\widehat{\alpha}$ on learning. At the beginning of learning, since MI has a greater influence than Δ, $\widehat{\alpha}$ is usually determined dependently on MI (left of Fig. 2). On the other hand, the critic network is trained such that the dependence of Z_s and Z_w is lowered. So, if only MI is used in the later stage of learning, useful candidates may be missed (top right). At this time, when Δ is added, gradients of useful candidates that MI could not consider can be used auxiliary (bottom right).

Entire Learning Procedure

Algorithm 1 describes the overview of CAF. First, the thresholds of arousal and valence, i.e., δ_a and δ_v are determined by a famous Otsu algorithm (Otsu 1979), and the decision threshold δ_r is calculated from δ_a and δ_v. Based on δ_r, two distribution of strong and weak emotions, i.e., $\mathbb{P}$ and $\mathbb{Q}$ are generated. Next, the sets of X_s and X_w sampled from $\mathbb{P}$ and $\mathbb{Q}$, respectively, are applied to compute $\mathcal{L}_c$ and $\mathcal{L}_{adv}$, and then θ_c and θ_e are updated. Here, d_θ represents the gradient of θ. Lastly, $\mathcal{L}_{AV}$ is optimized by the encoder network that adversarially learns the embedding representation of the critic network. Here, θ_r and θ_e are updated.

Further Analysis

Relationship between VDM and contrastive loss: $\mathcal{L}_c$ is defined from Eq. (3), Eq. (4), and $\widehat{\alpha}$.

$$\begin{aligned}\mathcal{L}_c = \sup_{f_{cont}\in\mathcal{F}} & \mathbb{E}_{(\mathbf{z}_i,\mathbf{z}_j)\sim\mathbb{P}} - D^2(\mathbf{z}_i,\mathbf{z}_j) \\ & - \frac{1}{2}\mathbb{E}_{(\mathbf{z}_i,\mathbf{z}_j)\sim(\mathbb{P}+\mathbb{Q})}\left(\widehat{\alpha} - D^2(\mathbf{z}_i,\mathbf{z}_j)\right)\end{aligned} \tag{6}$$

Here, the regularization term is omitted, and only the case where the hinge function of Eq. (4) is greater than 0 is considered. From the VDM point of view, $\mathcal{L}_c$ learns such that the difference between the first and second terms is maximized. If Eq. (6) is learned to minimize $\mathcal{L}_c$, it can be rewritten as follows:

$$\begin{aligned}\mathcal{L}_c &= \inf_{f_{cont}\in\mathcal{F}} \mathbb{E}_{(\mathbf{z}_i,\mathbf{z}_j)\sim\mathbb{P}} D^2\left(\mathbf{z}_i,\mathbf{z}_j\right) \\ &+ \frac{1}{2}\mathbb{E}_{(\mathbf{z}_i,\mathbf{z}_j)\sim(\mathbb{P}+\mathbb{Q})}\left(\widehat{\alpha} - D^2\left(\mathbf{z}_i,\mathbf{z}_j\right)\right) \\ &= \inf_{f_{cont}\in\mathcal{F}} \frac{1}{|\mathbb{P}|}\sum_{(\mathbf{z}_i,\mathbf{z}_j)\sim\mathbb{P}} D^2\left(\mathbf{z}_i,\mathbf{z}_j\right) \\ &+ \frac{1}{2\,|\mathbb{P}+\mathbb{Q}|}\sum_{(\mathbf{z}_i,\mathbf{z}_j)\sim(\mathbb{P}+\mathbb{Q})}\left(\widehat{\alpha} - D^2\left(\mathbf{z}_i,\mathbf{z}_j\right)\right)\end{aligned} \tag{7}$$

Monte Carlo estimation was applied to derive Eq. (7) in the form of the sum of individual samples. From the perspective of individual vectors, the first term of Eq. (7) is trained to increase the intra-class compactness of vectors within the same group, and the second term is learned to increase the inter-class variability of vectors between different groups.

Note that if the assumption of $f = f_{cont}$ is added in Eq. (3), the learning objectives of the lower bound of d_φ, i.e., VDM and contrastive loss become the same.

Analysis of adaptive margin: Let's analyze the adaptive margin $\widehat{\alpha}$. If the joint distribution of two random variables has the form of Gaussian distribution, the MI of $\widehat{\alpha}$ can be represented by $\mathrm{MI}(g(Z_s), g(Z_w)) = -\frac{1}{2}\log(1-\eta^2)$. Here, η indicates the correlation coefficient of $g(Z_s)$ and $g(Z_w)$ (Gel'Fand and Yaglom 1959). Δ of $\widehat{\alpha}$ is defined by $\Delta\,(N,K) = \log\left(\frac{N}{c}\right)^{\frac{1}{2K}}$. Here, K is the number of samples in the margin that is automatically determined by the empirical bias. The confidence parameter c is set to 0.05. Thus, the adaptive margin can be rearranged as follows:

$$\begin{aligned}\widehat{\alpha} &= -\frac{1}{2}\log(1-\eta^2) + \log\left(\frac{N}{c}\right)^{\frac{1}{2K}} \\ &= \log(1-\eta^2)^{-\frac{1}{2}}\left(\frac{N}{c}\right)^{\frac{1}{2K}}\end{aligned} \tag{8}$$

Eq. (8) can be interpreted as the product of correlation and adaptive scale. If the scale factor does not exist, $\widehat{\alpha}$ will approach zero when η^2 has a small value close to zero. This

Algorithm 1: Contrastive Adversarial Framework

Inputs: Parameters of encoder, critic, and FC layer $\theta_e, \theta_c, \theta_r$, learning rate ϵ, size of minibatch N;
while $\theta_e, \theta_c, \theta_r$ *converge* **do**
  Otsu threshold to decide δ_a and δ_v
  Calculate decision threshold $\delta_r = \sqrt{\delta_a^2 + \delta_v^2}$
  Decide two emotion distributions $\mathbb{Q}(i) = \left\{i \mid 0 \leq \sqrt{y_{a_i}^2 + y_{v_i}^2} < \delta_r\right\}$ and $\mathbb{P}(i) = \left\{i \mid \delta_r \leq \sqrt{y_{a_i}^2 + y_{v_i}^2} < 1\right\}$, $i = \{1, ..., N\}$;
  Sample $X_w = \{\mathbf{x}_{w_i}\}_{i=1}^{N_w}, \mathbf{x}_{w_i} = \mathbf{x}_i \sim \mathbb{Q}$
  Sample $X_s = \{\mathbf{x}_{s_i}\}_{i=N_w+1}^{N}, \mathbf{x}_{s_i} = \mathbf{x}_i \sim \mathbb{P}$;
  Feedforward $Z_{s/w} = (f_c \circ f_e)\left(X_{s/w}\right)$;
  $d_{\theta_c} \leftarrow \nabla_{\theta_c} \mathcal{L}_c\left(Z_s, Z_w\right)$ (Eq. (6))
  $\theta_c \leftarrow \theta_c + \epsilon \mathrm{ADAM}\left(\theta_c, d_{\theta_c}\right)$ **Freeze** θ_c;
  $d_{\theta_e} \leftarrow \nabla_{\theta_e} \mathcal{L}_{adv} = -\nabla_{\theta_e} \frac{1}{N_w} \sum_{i=1}^{N_w} (f_c \circ f_e)(\mathbf{x}_{w_i})$
  $\theta_e \leftarrow \theta_e - \epsilon \mathrm{ADAM}\left(\theta_e, d_{\theta_e}\right)$ **Unfreeze** θ_c;
  $(d_{\theta_e}, d_{\theta_r}) \leftarrow \nabla_{\theta_e} \nabla_{\theta_r} \mathcal{L}_{AV} = \nabla_{\theta_e} \nabla_{\theta_r} \frac{1}{N} \sum_{i=1}^{N} \{(f_r \circ f_e)(\mathbf{x}_i) - \mathbf{y}_i\}$
  $(\theta_e, \theta_r) \leftarrow (\theta_e, \theta_r) - \epsilon \mathrm{ADAM}\left(\theta_e, \theta_r, d_{\theta_e}, d_{\theta_r}\right)$;
end

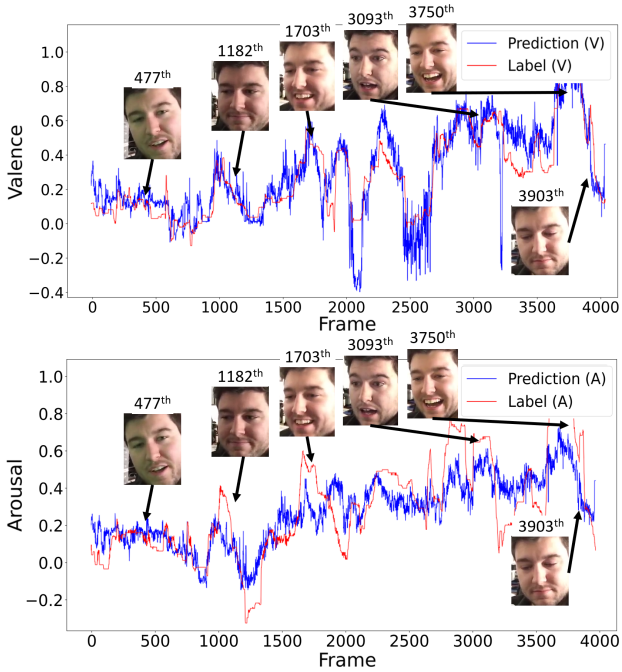

Figure 3: Qualitative analysis on Aff-Wild dataset.

makes learning unstable. Otherwise, even when η^2 becomes close to 0, $\hat{\alpha}$ has a non-zero value. That is, Δ not only increases the stability of similarity learning, but also adaptively considers vectors useful for learning.

Experiments

Datasets

AffectNet (Mollahosseini, Hasani, and Mahoor 2017) dataset consists of over a million images. Discrete categories as well as AV labels of about $440,000$ static images are annotated. This is classified as a wild dataset because of its complex background, illumination, and point of view.

AFEW-VA (Kossaifi et al. 2017) dataset is derived from the AFEW (Dhall et al. 2016) dataset for the EMOTIW challenge. This dataset is constructed in video sequence units, but AV labels are annotated in a frame basis.

Aff-Wild (Zafeiriou et al. 2017) dataset consists of about 300 videos of various experimental participants watching TV shows and movies. The test dataset is not released. So, a part of the training dataset is randomly selected and used as an evaluation dataset in this paper, same as (Hasani, Negi, and Mahoor 2020).

Training Configurations

Implementation details: All experiments were performed on the Intel Xeon CPU and GeForce GTX 1080 TI, with five training sessions per experiment [1]. AlexNet and ResNet18 were used as the encoder network to learn facial feature information from scratch. The details of network structure are described in the supplementary material. Encoder, critic, and FC layers were optimized through the learning rate of Adam (Kingma and Ba 2014) optimizer with 1e-4. The minibatch size of AlexNet and ResNet18 were set to 256 and 128, respectively. The parameters were updated through 50,000 iterations for the AffectNet and AFEW-VA datasets, and 100,000 iterations for the Aff-Wild dataset. We reduced the learning rate by 0.8 times every 10k iterations. The emotion grouping process using Otsu thresholding and the critic network were used only in the training process, and only the encoder and FC layers operated in the inference process (see Fig. 1). On the other hand, in the experiments using the DML dataset, the learning configurations of the base methods (Wu et al. 2017) were referenced as they are.

Evaluation metrics: In this paper, root mean squared error (RMSE) is a metric that quantifies the difference between a label and the estimate: $RMSE(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\mathbb{E}\left(\mathbf{y} - \hat{\mathbf{y}}\right)^2}$. Here, $\mathbf{y}$ and $\hat{\mathbf{y}}$ indicate the label and the estimate. Sign agreement (SAGR) metric measures the degree of positive or negative emotions overall, which is generally difficult to judge with RMSE: $SAGR(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^{N} \Gamma\left(\mathrm{sign}(\mathbf{y}_i), \mathrm{sign}(\hat{\mathbf{y}}_i)\right)$. Here, Γ is a Kronecker delta function that outputs 1 if two inputs are the same, and 0 otherwise. Pearson correlation coefficient (PCC) is used as a metric that quantifies the similarity between labels and estimates. PCC overcomes the disadvantage of RMSE, which cannot take into account the proportion of outlier samples: $PCC = \frac{COV(\mathbf{y}, \hat{\mathbf{y}})}{\sigma_{\mathbf{y}} \sigma_{\hat{\mathbf{y}}}} = \frac{\mathbb{E}(\mathbf{y} - \mu_{\mathbf{y}})(\hat{\mathbf{y}} - \mu_{\hat{\mathbf{y}}})}{\sigma_{\mathbf{y}} \sigma_{\hat{\mathbf{y}}}}$. Here, $COV(\cdot, \cdot)$ is the covariance, $\mu_{\mathbf{y}}$ and $\sigma_{\mathbf{y}}$ indicate the mean and standard

[1] Software is available at https://github.com/kdhht2334/Contrastive-Adversarial-Learning-FER

Methods	Backbone	Params.	RMSE		PCC		CCC	
			(V)	(A)	(V)	(A)	(V)	(A)
(Mollahosseini, Hasani, and Mahoor 2017)	AlexNet	61M	0.37	0.41	0.66	0.54	0.60	0.34
(Jang, Gunes, and Patras 2019)	SSD w/ VGG16	-	0.44	0.39	0.58	0.50	0.57	0.47
(Kollias et al. 2018)	VGG16	-	0.37	0.39	0.66	0.55	0.62	0.54
(Barros, Parisi, and Wermter 2019)	AlexNet	-	-	-	-	-	0.67	0.38
(Kossaifi et al. 2020)	ResNet18	-	0.35	0.32	0.71	0.63	0.71	0.63
(Hasani, Negi, and Mahoor 2020)	ResNeXt50	3.1M	0.2668	0.2482	0.78	0.86	0.74	0.85
Ours	ResNet18	11M	**0.2186**	**0.1873**	**0.86**	0.85	**0.83**	0.84
	AlexNet (tuned)	3.6M	0.2216	0.1916	0.81	**0.86**	0.80	**0.85**

Table 1: Results on the AffectNet dataset. '(V)' and '(A)' represent valence and arousal, respectively. AlexNet (tuned) is a reduced model of FC layers of AlexNet.

Case	Methods	RMSE		SAGR		PCC		CCC	
		(V)	(A)	(V)	(A)	(V)	(A)	(V)	(A)
Static	(Kossaifi et al. 2017)	0.27	0.23	-	-	0.41	0.45	-	-
	(Mitenkova et al. 2019)	0.40	0.41	-	-	0.33	0.42	0.33	0.40
	(Kossaifi et al. 2020)	0.24	0.24	0.64	0.77	0.55	0.57	0.55	0.52
	Ours (ResNet18)	**0.17**	**0.18**	**0.68**	**0.87**	**0.67**	0.60	**0.59**	0.54
	Ours (AlexNet (tuned))	0.20	0.20	0.66	0.83	**0.67**	**0.63**	0.58	**0.57**
Temporal	ResNet18-3D	0.26	0.22	0.56	0.77	0.19	0.33	0.17	0.29
	ResNet18-(2+1)D	0.31	0.29	0.50	0.73	0.17	0.33	0.16	0.20
	(Kollias et al. 2019)	-	-	-	-	0.51	0.58	0.52	0.56
	(Kossaifi et al. 2020)-scratch	0.28	0.19	0.53	0.75	0.12	0.23	0.11	0.15
	(Kossaifi et al. 2020)-trans.	0.20	0.21	0.67	0.79	0.64	0.62	0.57	0.56

Table 2: Results on the AFEW-VA dataset.

deviation of $\mathbf{y}$. Concordance correlation coefficient (CCC) is a metric that quantifies similarity by fusing PCC and the statistics of labels and estimates: $CCC = \frac{2\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}PCC(\mathbf{y},\hat{\mathbf{y}})}{\sigma_{\mathbf{y}}^2+\sigma_{\hat{\mathbf{y}}}^2+(\mu_{\mathbf{y}}-\mu_{\hat{\mathbf{y}}})^2}$.

In order to learn AV domain emotions effectively, minimization of RMSE and maximization of PCC/CCC should be performed at the same time (Kossaifi et al. 2020). Therefore, PCC and CCC terms are added to the final loss function of the proposed FER: $\mathcal{L}_{AV} = \mathcal{L}_{RMSE} + \epsilon\left(\mathcal{L}_{PCC} + \mathcal{L}_{CCC}\right)$. Here, $\mathcal{L}_{PCC} = 1 - \frac{PCC_a+PCC_v}{2}$, $\mathcal{L}_{CCC} = 1 - \frac{CCC_a+CCC_v}{2}$, and $\epsilon = 0.5$.

Performance Comparison

Quantitative evaluation: Table 1 shows the experimental results for the AffectNet dataset. The proposed method is better than the latest algorithms such as FHC (Kossaifi et al. 2020) and BreG-NeXt (Hasani, Negi, and Mahoor 2020). For the ResNet18 backbone, the proposed method showed 0.05 lower RMSE than BreG-NeXt. Because of the direct pipeline structure between the input and label supervision, BreG-NeXt was negatively affected by some features other than facial expressions. In addition, comparing the numerical results of (Barros, Parisi, and Wermter 2019) where a person-independent FER was first proposed, the proposed method showed higher CCC(V) by up to 0.16. This demonstrates the superiority of the proposed method in terms of the ability to recognize emotions independently of persons.

Table 2 is the results for the AFEW-VA dataset. In this experiment, various static methods as well as temporal methods were compared. Note that although the proposed method never uses temporal information, it has achieved superior RMSE(V) and PCC(V) than the SOTA temporal method, i.e., FHC. This proves that FER focusing on emotional changes like the proposed method is very competitive.

Table 3 depicts the generalization performance of the proposed method through an experiment using the Aff-Wild dataset. In all evaluation metrics except SAGR(V), the proposed method outperformed BreG-NeXt. For example, in terms of CCC(A), the proposed method showed improvement of 0.25 over BreG-NeXt.

Qualitative evaluation: Figure 3 plots the predicted label values on a frame basis. For this experiment, the evaluation dataset of the Aff-Wild dataset was employed. The proposed method catches up well with the ground-truth label flow having significant fluctuations. Note that the estimated AV values match the actual facial expressions of the subject very accurately. In addition, the visualization results using T-SNE (Maaten and Hinton 2008) demonstrate that the proposed method achieved person-independent FER and the proposed discriminative learning was effective. This additional results are described in the supplementary material.

Ablation Study

Impact of adversarial learning methods: In order to analyze the influence of adversarial learning on the overall performance, various MMD and IPM-based techniques (GramNet , SphereGAN, FisherGAN) (Srivastava et al. 2019; Park and Kwon 2020; Mroueh and Sercu 2017) and the proposed CAF were compared in Table 4. Adversarial learning in Table 4 indicates a method that replaces Eq. (6) in Algorithm

Methods	RMSE		SAGR		PCC		CCC	
	(V)	(A)	(V)	(A)	(V)	(A)	(V)	(A)
(Hasani and Mahoor 2017)	0.27	0.36	0.57	0.74	0.44	0.26	0.36	0.19
(Hasani, Negi, and Mahoor 2020)	0.26	0.31	**0.77**	0.75	0.42	0.40	0.37	0.31
(Deng et al. 2019)*	-	-	-	-	-	-	**0.58**	0.52
Ours (ResNet18)	**0.22**	**0.20**	0.70	0.76	**0.57**	**0.57**	0.55	**0.56**
Ours (AlexNet (tuned))	0.24	0.21	0.68	**0.78**	0.55	**0.57**	0.54	**0.56**

Table 3: Results on the Aff-Wild dataset.* indicates the results on Aff-Wild $test$ dataset with ResNet50 backbone.

Methods	Backbone	Critic	Params.	RMSE		SAGR		CCC	
				(V)	(A)	(V)	(A)	(V)	(A)
BreG-NeXt	ResNeXt50	-	3.1M	0.2668	0.2482	0.77	0.82	0.74	**0.85**
Adversarial learning	ResNet18	-	10M	0.3687	0.3584	0.74	0.76	0.64	0.55
		FisherGAN	11M	0.2221	0.1994	0.78	0.82	0.81	0.82
		SphereGAN		0.2467	0.2152	0.77	0.78	0.77	0.80
		GramNet		0.2251	0.1985	**0.80**	0.82	0.78	0.82
	AlexNet (tuned)	-	3.0M	0.3712	0.3747	0.68	0.70	0.62	0.38
		FisherGAN	3.6M	0.2437	0.2014	0.75	0.80	0.75	0.78
		SphereGAN		0.2687	0.2212	0.72	0.78	0.72	0.79
		GramNet		0.2360	0.2046	0.78	0.82	0.76	0.80
Ours w/ Δ	ResNet18	CAF	11M	**0.2186**	**0.1873**	**0.80**	**0.84**	**0.83**	0.84
	AlexNet (tuned)		3.6M	0.2216	0.1916	0.79	**0.84**	0.80	**0.85**
Ours w/ constant κ	ResNet18		11M	0.2289	0.2082	0.78	0.82	0.76	0.80
	AlexNet (tuned)		3.6M	0.2422	0.2158	0.77	0.81	0.74	0.80
Ours w/ linear model κ'	ResNet18		11M	0.2201	0.1982	**0.80**	**0.84**	0.80	0.84
	AlexNet (tuned)		3.6M	0.2255	0.1994	0.79	0.82	0.79	0.84

Table 4: Results of ablation study experiments using AffectNet dataset.

1 with the existing MMD or IPM. In this case, the backbone of ResNet18 without critic showed 0.1 higher RMSE than BreG-NeXt. However, when adversarial learning with the GramNet structure added is applied, the proposed method showed 0.04 lower RMSE(V) and 0.03 higher SAGR(V) than BreG-NeXt. This proves that adversarial learning plays a big role in improving FER performance.

In addition, ours with Δ showed superior performance than the other adversarial learning techniques. This indirectly proves that the VDM-based contrastive loss function, which learns the similarity from the perspective of individual samples, enables more effective learning than the IPM that learns the overall distribution difference.

Further analysis of adaptive margin: The effect of the confidence interval Δ of the adaptive margin on the learning of the critic network was analyzed. First, Δ was replaced with an appropriate constant κ $(= 0.2)$, that is the best constant. As in Table 4, RMSE increased compared to GramNet when κ was used. This performance degradation is because κ cannot take into account the density and statistical distribution of vectors in latent space at all.

Next, we examined a linear model $\kappa' = 0.1 + (\text{epoch}) \times 5\text{e-}3$ by referring to the linearly proportional characteristic of Δ. Here, epoch represents the number of times that training has been completed for the entire dataset. As in Table 4, κ' was better than κ. However, its RMSE was still higher than Δ. Thus, considering the distribution of the latent space, useful samples can be found.

Grouping of emotion samples: To verify that Otsu algorithm is nearly optimal for grouping emotional samples, we calculated k cluster centers per domain using k-means clustering (Kanungo et al. 2002), and averaged them to determine the decision threshold. We compared the experimental results for k of 2, 3, and 5 with the results of 'ours with Δ' in Table 4. We observed a marginal difference of about 0.05 in RMSE, and only a slight difference of about 0.02 in CCC. Although the Otsu algorithm is a non-learning approach, it can cluster emotion samples effectively enough for the proposed method.

DML with adaptive margin: Since most DML techniques learn samples based on margin, the role of margin is very important. So, we applied the proposed adaptive margin to recent DML techniques and analyzed how much it affects to reflect the statistical distribution of latent space. As a result, in recent DML techniques, the proposed scheme has achieved the SOTA performance. Please refer to the supplementary material for detailed experimental results.

Conclusion

While conventional methods tried to improve the performance of the AV domain FER through the improvement of CNN, the proposed method focused on learning the degree of emotional change through adversarial learning. This study will be a great inspiration to existing studies that have concentrated only on the network structure. Also, this study is valuable in that it proposed a novel contrastive loss function and proved its effectiveness in deep metric learning as well as adversarial emotion learning.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) [2020-0-01389, Artificial Intelligence Convergence Research Center (Inha University)] and Industrial Technology Innovation Program through the Ministry of Trade, Industry, and Energy (MI, Korea) [Development of Human-Friendly Human-Robot Interaction Technologies Using Human Internal Emotional States] under Grant 10073154.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. volume 70 of *Proceedings of Machine Learning Research*, 214–223. International Convention Centre, Sydney, Australia: PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Balconi, M.; Vanutelli, M. E.; and Finocchiaro, R. 2014. Multilevel analysis of facial expressions of emotion and script: self-report (arousal and valence) and psychophysiological correlates. *Behavioral and Brain Functions* 10(1): 32.

Balsubramani, A.; Dasgupta, S.; Moran, S.; et al. 2019. An adaptive nearest neighbor rule for classification. In *Advances in Neural Information Processing Systems*, 7579–7588.

Barros, P.; Parisi, G.; and Wermter, S. 2019. A personalized affective memory model for improving emotion recognition. In *International Conference on Machine Learning*, 485–494.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062* .

Cai, J.; Meng, Z.; Khan, A. S.; Li, Z.; O'Reilly, J.; and Tong, Y. 2019. Identity-free facial expression recognition using conditional generative adversarial network. *arXiv preprint arXiv:1903.08051* .

Deng, D.; Chen, Z.; Zhou, Y.; and Shi, B. 2019. MIMAMO Net: Integrating Micro-and Macro-motion for Video Emotion Recognition. *arXiv preprint arXiv:1911.09784* .

Dhall, A.; Goecke, R.; Joshi, J.; Hoey, J.; and Gedeon, T. 2016. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM international conference on multimodal interaction*, 427–432.

Gel'Fand, I.; and Yaglom, A. 1959. ABouTA RANDoM FUNCTION contained IN ANOTHER SUCH FUNCTION". *Eleven Papers on Analysis, Probability and Topology* 12: 199.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13(1): 723–773.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, 1735–1742. IEEE.

Harmon-Jones, E.; Harmon-Jones, C.; Amodio, D. M.; and Gable, P. A. 2011. Attitudes toward emotions. *Journal of personality and social psychology* 101(6): 1332.

Hasani, B.; and Mahoor, M. H. 2017. 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*, 9–16.

Hasani, B.; Negi, P. S.; and Mahoor, M. 2020. BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient. *IEEE Transactions on Affective Computing* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hiriart-Urruty, J.-B.; and Lemaréchal, C. 2012. *Fundamentals of convex analysis*. Springer Science & Business Media.

Jang, Y.; Gunes, H.; and Patras, I. 2019. Registration-free Face-SSD: Single shot analysis of smiles, facial attributes, and affect in the wild. *Computer Vision and Image Understanding* 182: 17–29.

Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; and Wu, A. Y. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence* 24(7): 881–892.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kollias, D.; Cheng, S.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2018. Generating faces for affect analysis. *arXiv preprint arXiv:1811.05027* 16.

Kollias, D.; Tzirakis, P.; Nicolaou, M. A.; Papaioannou, A.; Zhao, G.; Schuller, B.; Kotsia, I.; and Zafeiriou, S. 2019. 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.

Kossaifi, J.; Toisoul, A.; Bulat, A.; Panagakis, Y.; Hospedales, T. M.; and Pantic, M. 2020. 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*, 6060–6069.

Kossaifi, J.; Tzimiropoulos, G.; Todorovic, S.; and Pantic, M. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65: 23–36.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Mitenkova, A.; Kossaifi, J.; Panagakis, Y.; and Pantic, M. 2019. Valence and arousal estimation in-the-wild with tensor methods. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–7. IEEE.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10(1): 18–31.

Mroueh, Y.; and Sercu, T. 2017. Fisher gan. In *Advances in Neural Information Processing Systems*, 2513–2523.

Mroueh, Y.; Sercu, T.; and Goel, V. 2017. Mcgan: Mean and covariance feature matching gan. *arXiv preprint arXiv:1702.08398* .

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2005. On divergences, surrogate loss functions and decentralized detection. *arXiv preprint math.ST/0510521* .

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* 56(11): 5847–5861.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, 271–279.

Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* 9(1): 62–66.

Park, S. W.; and Kwon, J. 2020. SphereGAN: Sphere Generative Adversarial Network Based on Geometric Moment Matching and its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Russell, J. A. 2017. Cross-cultural similarities and differences in affective processing and expression. In *Emotions and affect in human factors and human-computer interaction*, 123–141. Elsevier.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Srivastava, A.; Xu, K.; Gutmann, M. U.; and Sutton, C. 2019. Generative Ratio Matching Networks. In *International Conference on Learning Representations*.

Vielzeuf, V.; Pateux, S.; and Jurie, F. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 569–576.

Wu, C.-Y.; Manmatha, R.; Smola, A. J.; and Krahenbuhl, P. 2017. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2840–2848.

Yang, H.; Ciftci, U.; and Yin, L. 2018. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2168–2177.

Zafeiriou, S.; Kollias, D.; Nicolaou, M. A.; Papaioannou, A.; Zhao, G.; and Kotsia, I. 2017. Aff-wild: Valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 34–41.

Zeng, J.; Shan, S.; and Chen, X. 2018. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, 222–237.

Zhang, F.; Zhang, T.; Mao, Q.; and Xu, C. 2018. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3359–3368.