IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 9, 2020, accepted August 10, 2020, date of publication August 24, 2020, date of current version September 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018738

Toward Unbiased Facial Expression Recognition in the Wild via Cross-Dataset Adaptation

BYUNGOK HAN[1], WOO-HAN YUN[1], JANG-HEE YOO[1], (Senior Member, IEEE), AND WON HWA KIM[2], (Member, IEEE)

[1]Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, South Korea
[2]Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76010, USA

Corresponding author: Won Hwa Kim (won.kim@uta.edu)

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00330, Development of AI Technology for Early Screening of Infant/Child Autism Spectrum Disorders based on Cognition of the Psychological Behavior and Response).

ABSTRACT Despite various success in computer vision with facial images (e.g., face detection, recognition, and generation), facial expression recognition is still a challenging problem yet to be solved. This is because of simple but fundamental bottlenecks: 1) no global agreement on different facial expressions, 2) significant dataset biases that prevent cross-dataset analysis for a large-scale study, and 3) high class imbalance in in-the-wild datasets that causes inconsistency in predicting expressions in images using a machine learning algorithm. To tackle these issues, we propose a novel Deep Learning approach via adaptive cross-dataset scheme. We combine multiple in-the-wild datasets to secure sufficient training samples while minimizing dataset bias using ideas of reversal gradients to retain generality. For this, we introduce a flexible objective function that can control for skewed label distributions in the dataset. Incorporating these ideas, together with the ResNet pipeline as a backbone, we carried extensive experiments to validate our ideas using three independent in-the-wild facial expression datasets, which first confirmed bias from different datasets and yielded improved performance on facial expression recognition using the multi-site dataset.

INDEX TERMS Cross-dataset bias, deep learning, domain adaptation, facial expression recognition, in-the-wild dataset.

I. INTRODUCTION

Face is a primary means to transfer information not only among humans but it is an effective tool for communication between humans and machines as well. In this regard, analyses of faces using images have been adopted for fundamental researches in various areas such as neuroscience [1], psychology [2], human-computer-interaction [3], etc. It is indisputable that computer vision methods are the driving forces of such researches; there is a rich history of works in vision with face detection [4], [5], face recognition [6], [7], 3D face construction [8], [9], facial image generation [10], [11] as well as substantial extensions in security such as personal identification [12], face spoofing and anti-spoofing [13], [14]. Recent works with Deep Learning (DL) demonstrate remarkable advances in these applications by combining a flexible neural network that can train a generalized model and large-scale training samples that recently became available, e.g., [15], [16]. The prediction performance of these algorithms is up to human-level precision [16] and these methods are deployed on many commercial devices with cameras.

The visual information from a face is mainly delivered by facial expressions. However, despite a rich body of successful works with facial images in machine learning and computer vision, facial expression recognition still remains a challenging task that is yet to be solved. First of all, we face a lack of data with reliable annotations to train a data-driven machine learning model. This problem stems from two issues: 1) personal facial expressions are driven from subjective emotions that are difficult to assign objective labels [17], and 2) there exist substantial ambiguity between emotion classes [18]. That is, there is a severe diversity in expression caused by personal and cultural differences that lead to large intra-class variation and small inter-class variation. For example, the same emotional expression shown

The associate editor coordinating the review of this manuscript and approving it for publication was Paolo Remagnino.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

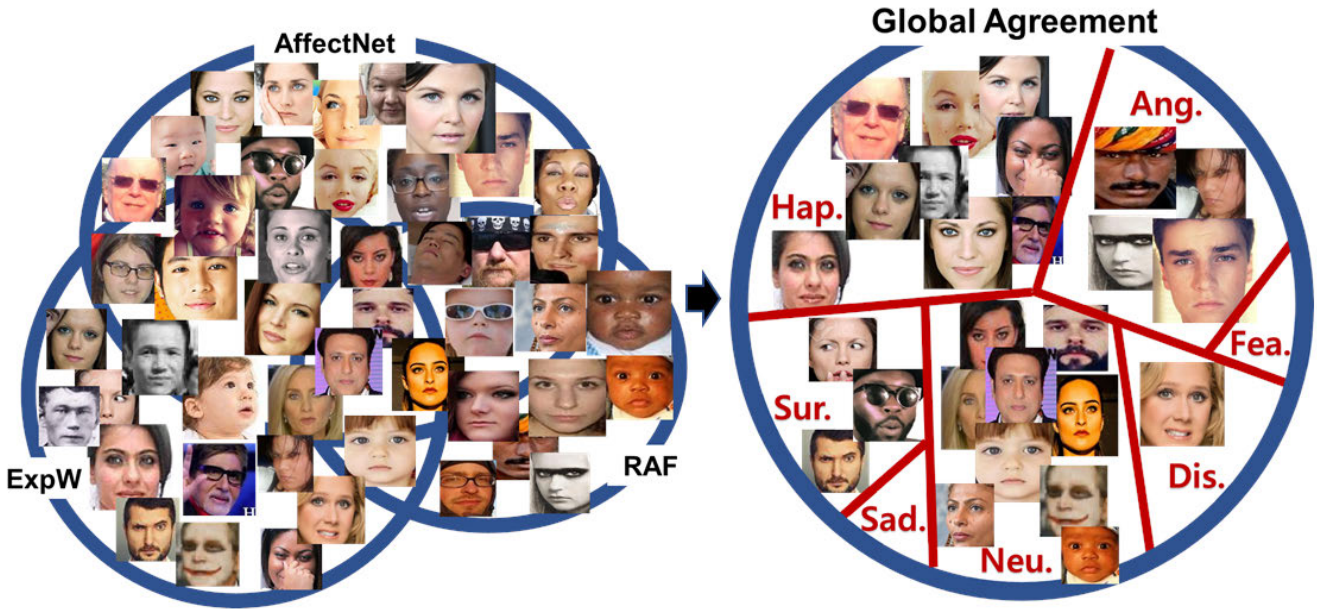


FIGURE 1. Facial expression images of the same class (Neutral) across different datasets. Although all images are annotated as the neutral class, actual facial expressions may be inconsistent across the three datasets.

in an image can be categorized differently in a different dataset and vise versa for different emotions (See Fig. 1). Also, a mixture of emotions, which often arise in various real facial expressions, makes the expression recognition problem even more challenging. Such issues are clearly distinguished from recent popular datasets such as ImageNet [19] and MS-Celeb-1M [20], whose annotation and identity information are creditable.

Notably, there have been several tries to tackle the facial expression recognition problem in a data-driven way. However, individual dataset acquired at different sites often includes substantial dataset bias caused by local data collection protocol and the aforementioned diversity in expressions. These biases include *selection bias* caused at the image searching stage using limited key words or the participants recruiting stage, *capturing bias* caused by the data collection environment, and *annotation bias* caused at the image labeling stage [21]. In the end, regardless of classification algorithms, these (potentially large) biases make it difficult for the algorithms to agree on the same expression coming from different images. Datasets collected in the past two decades are mostly in-the-lab setting where their participants *acted* to make specific expressions or to induce to make spontaneous expressions based on the psychological ground in restricted lab environments, which eventually leads to exaggerated expressions [22]. Some of the recent datasets broke away from restrictions from the labs and collected a large number of online facial expression images based on a set of keywords. These crawled images were then annotated to constitute *in-the-wild* datasets [22]–[24].

One approach to reducing the annotation bias is to have a diverse group of annotators to soft-label the data with probability [25], however, it requires substantial resources and efforts. The simplest way would be to perform cross-dataset generalization: to combine data from multi-sites to secure a sufficient number of samples [26], [27]. This is conceptually easy but reducing annotation bias across datasets remains a challenging task. There have been several tries [28]–[31], but these works often lack explanation on dataset invariance, suffer from complex architecture with label combination, overfit to a single dataset, and cannot control for selection and annotation biases. Moreover, even if we have enough data, many facial expression datasets have class imbalance problem that affects downstream classification, which is common in vision datasets [19]. The class imbalance occurs more frequently in the in-the-wild datasets since the frequencies of each expression depends on the emotion type. For example, facial expressions are typically categorized by 6 basic emotions (i.e., Happiness, Surprise, Disgust, Sadness, Fear and Anger) and 1 neutrality, and some of these emotions are expressed more often than others. Such natural phenomena affect the skewed distribution of data labels in the wild; $\sim$70% of class labels in many in-the-wild datasets belong to either Happiness or Neutral.

In this regard, we tackle the issues addressed above by proposing an adaptive cross-dataset framework let us combine multiple datasets to constitute a larger-scale multi-site dataset with minimal dataset bias. The framework is a variant of domain adaptation methods; unlike other tries to reduce differences between source and target domains [32], the key idea is to rather *minimize bias between datasets* to boost facial expression recognition performance with a global multi-site dataset. We deal with two of the biases introduced above: 1) to reduce selection bias, we train our model such that it minimizes differences in distributions of image features extracted from different datasets, and 2) to reduce annotation bias, we add a label extractor in our pipeline that generates pseudo emotion labels used as image feature/label pairs to train our

model such that it does not distinguish different datasets. Finally, in order to tackle the class imbalance issue that typically arises in many in-the-wild datasets, we introduce an *adaptive cross entropy* as an objective function that assigns different weights to class-wise cross entropy according to a statistical property of each class and training precision during optimization.

Our work throughout this article suggests the following **contributions**: 1) we successfully combine multiple in-the-wild facial expression datasets to provide a sufficient number of samples for training an machine learning (ML) model by minimizing biases between datasets, 2) we introduce an adaptive cross entropy scheme to work around the skewed class distribution of facial expressions in the dataset, 3) we demonstrate extensive empirical experimental results identifying dataset bias and validating the performance of our framework on facial expression recognition using the aggregated data from multiple in-the-wild datasets. The results show that our framework is able to accurately classify various facial expressions and performs better than other state-of-the-art baseline methods.

II. PROPOSED METHOD

We tackle the two major issues in facial expression recognition with multi-site datasets with in-the-wild images: 1) dataset bias, and 2) class imbalance problems. We propose a cross-dataset adaptation (CDA) scheme to address selection and annotation biases from different datasets using separate feature extractor and pseudo-label extractor, and adaptively control classification error from cross-entropy to mitigate the class imbalance problem. The details are given below.

A. CROSS-DATASET ADAPTATION

Combining multiple datasets from multi-sites is a common approach to secure a sufficient number of meaningful samples. However, the bias introduced from different datasets (i.e., cross-dataset bias) must be minimized to improve the generalization capability of an ML algorithm trained by the multi-site dataset. For facial expression datasets, one cause of such bias may be regional/cultural specific annotation resulting in inconsistent labels between different datasets. This means simply gathering and training with facial images and their labels can lead to harmful effects on generalization, especially when the datasets have very different characteristics or include conflicting annotations on the same expression.

In this scenario, our idea to control for the cross-dataset bias is to construct features robust to dataset type and utilize a pseudo label for emotion to reversely train on dataset label. We design two components: 1) a feature extractor, which learns image features using a Residual Network (ResNet) that reduces specificity between different datasets while increasing discriminant ability for emotion classes, and 2) an emotion label extractor using a Convolution Neural Network (CNN) that is used to reduce annotation inconsistency between datasets. We utilize a dataset classifier with a Gradient Reversal Layer (GRL) [32], which can be viewed as a control tower to minimize biases between datasets. The GRL *reversely* trains both the feature extractor and the label extractor *not to distinguish* different datasets while training the emotion classifier to accurately predict emotions based on the trained features. The overall architecture of our proposed method with CDA is shown in Figure 2. There are multiple emotion classifiers (three classifiers for our experiment), i.e., $e = (e_1, e_2, e_3)$, that are assigned to individual dataset within the full multi-site dataset. The details of our approach are described below.

Given input images $\mathbf{x}$ and their labels $\mathbf{y}$, we constitute a feature extractor $f(\mathbf{x}; \theta_\mathbf{f})$ with parameters $\theta_\mathbf{f}$ and a label extractor $g(\mathbf{y}; \theta_\mathbf{g})$ with parameters $\theta_\mathbf{g}$. Then, a CDA component is defined as $CDA(\mathbf{x}, \mathbf{y}; \theta_\mathbf{f}, \theta_\mathbf{g}, \theta_\mathbf{e}, \theta_\mathbf{d})$ that combines an emotion classifier $e(f(\mathbf{x}); \theta_\mathbf{e})$ and a dataset classifier $d(f(\mathbf{x}), g(\mathbf{y}); \theta_\mathbf{d})$, where θ_e and θ_d are trainable parameters in the emotion classifier and dataset classifier respectively. The loss function L_e for emotion classifier is defined as

$$L_e(\mathbf{x}, \mathbf{y}; \theta_\mathbf{f}, \theta_\mathbf{e}, \theta_\mathbf{g}) = L(e(f(\mathbf{x})), g(\mathbf{y}); \theta_\mathbf{f}, \theta_\mathbf{e}) + L_2(g(\mathbf{y}), \mathbf{y}; \theta_\mathbf{g}) \quad (1)$$

where, L is an error between an output from the emotion classifier $e(f(\mathbf{x}))$ and a pseudo label from the label extractor $g(\mathbf{y})$, and L_2 is a ℓ_2-regularizer for the weights of the label extractor $\theta_\mathbf{g}$ to make the pseudo label stable. The loss function L_d for dataset classifier is formulated with dataset label $\mathbf{z}$ and parameters $\theta_\mathbf{f}$, $\theta_\mathbf{g}$, $\theta_\mathbf{d}$ as

$$L_d(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta_\mathbf{f}, \theta_\mathbf{g}, \theta_\mathbf{d}) = L(d(f(\mathbf{x}), g(\mathbf{y})), \mathbf{z}; \theta_\mathbf{f}, \theta_\mathbf{g}, \theta_\mathbf{d}) \quad (2)$$

that takes $(f(\mathbf{x}), g(\mathbf{y}))$ at the same time to train the dataset classifier d. Combining these two losses L_e and L_d, our *multi-task loss* E for CDA becomes

$$E(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta_\mathbf{f}, \theta_\mathbf{g}, \theta_\mathbf{e}, \theta_\mathbf{d}) = L_e(\mathbf{x}, \mathbf{y}; \theta_\mathbf{f}, \theta_\mathbf{e}, \theta_\mathbf{g}) + L_d(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta_\mathbf{f}, \theta_\mathbf{g}, \theta_\mathbf{d}) \quad (3)$$

and (3) is further reformulated with the GRL at the backpropagation stage as:

$$E(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta_\mathbf{f}, \theta_\mathbf{g}, \theta_\mathbf{e}, \theta_\mathbf{d}) = L_e(\mathbf{x}, \mathbf{y}; \theta_\mathbf{f}, \theta_\mathbf{e}, \theta_\mathbf{g}) - \lambda_1 L_d(f(\mathbf{x}); \theta_\mathbf{f}, \theta_\mathbf{d}) - \lambda_2 L_d(g(\mathbf{y}); \theta_\mathbf{g}, \theta_\mathbf{d}) \quad (4)$$

where λ_1 and λ_2 are user parameters to balance effects of the feature extractor and the label extractor respectively from the dataset classifier. Minimizing the (4) jointly trains a DL model such that it correctly classifies different facial expressions in the images, and learns features from images and pseudo labels where intra-dataset variation are maximized simultaneously. The training is performed via backpropagation with partial derivatives given below:

$$\theta_\mathbf{f} \leftarrow \theta_\mathbf{f} - \mu \left(\frac{\partial L_e}{\partial \theta_\mathbf{f}} - \lambda_1 \frac{\partial L_d}{\partial \theta_\mathbf{f}} \right) \quad (5)$$

$$\theta_\mathbf{g} \leftarrow \theta_\mathbf{g} - \mu \left(\frac{\partial L_2}{\partial \theta_\mathbf{g}} - \lambda_2 \frac{\partial L_d}{\partial \theta_\mathbf{g}} \right) \quad (6)$$

$$\theta_\mathbf{e} \leftarrow \theta_\mathbf{e} - \mu \frac{\partial L_e}{\partial \theta_\mathbf{e}}, \ \theta_\mathbf{d} \leftarrow \theta_\mathbf{d} - \mu \frac{\partial L_d}{\partial \theta_\mathbf{d}} \quad (7)$$

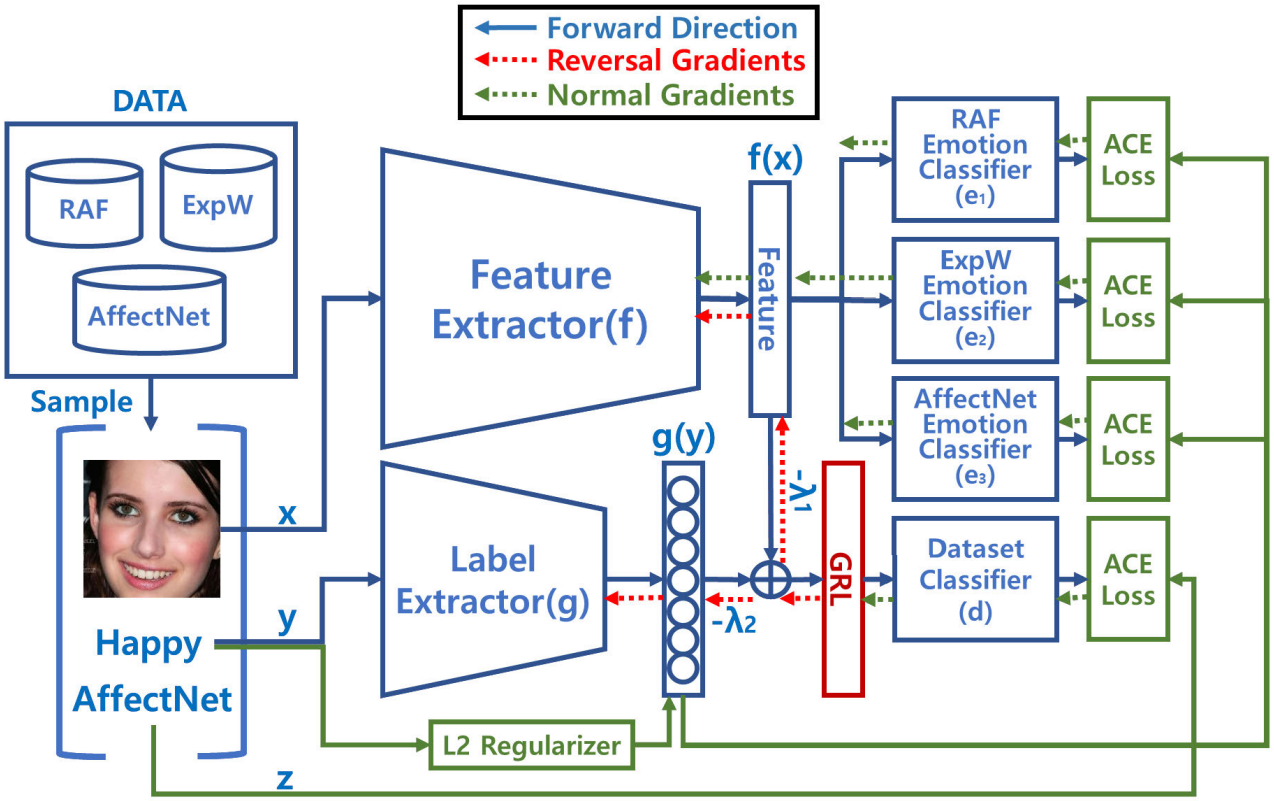


FIGURE 2. The overall architecture of our framework. Features derived from facial images x and altered labels (psuedo labels $g(y)$) from ground truths y are inputted to separate emotion classifiers e_1, e_2, e_3 and dataset classifier d to minimize emotion classification error and dataset bias based on the proposed Adaptive Cross Entropy (ACE) Loss.

where, μ is a learning rate of the overall DL structure. Notice that $\theta_{\mathbf{f}}$ is jointly trained by both L_e and L_d to satisfy both conditions for emotion classifier and dataset classifier. The $\theta_{\mathbf{g}}$ is trained by L_2 and L_d; specifically, the update with respect to L_d is achieved by implementing the GRL. The gradients of L_d are applied to $\theta_{\mathbf{f}}$ and $\theta_{\mathbf{g}}$ in the opposite direction, so that it degrades precision of dataset classifier and consider all datasets as a single dataset.

B. ADAPTIVE CROSS ENTROPY

Collecting a sufficient amount of samples for each emotion category is difficult by the nature of domains where the data are collected. Many datasets from research labs, i.e., in-the-lab datasets, collect posed facial expression images in a constrained laboratory environment (i.e., consistent pose, angle, illumination, etc.). These datasets usually have equally distributed samples for facial expression categories from ideal environments. Although the performance of ML models with in-the-lab datasets may be reasonable, these datasets cannot be generalized to the real-world environment since the data were acquired in restricted settings. This is because the models trained based on in-the-lab datasets cannot be adequately extended to real-life conditions, where the majority of pictures are taken in the wild. This can be easily verified by testing a pre-trained model (using in-the-lab data) with in-the-wild datasets, which causes a significant performance drop in a real-world scenario for facial expression classification [33].

We, therefore, must turn to in-the-wild datasets that provide better generality to a trainable model. Unfortunately, in a typical in-the-wild dataset, there exists a significant class imbalance problem. That is, the majority of images in the dataset are categorized as Happiness and Neutrality classes, whereas Fear, Disgust, Surprise, Sadness, and Anger classes have much fewer samples due to the inherent properties of emotional status. A simple solution to such a class imbalance problem is to balance the number of samples in all classes based on the sample size of a minority class. However, this may lead to a considerable reduction in overall data volume, which can significantly drop classification performance, especially with DL algorithms.

In such a scenario, the Adaptive Cross Entropy (ACE) proposed in this article is designed to reflect the characteristics of a dataset by constructing a loss function based on the distribution and precision of prediction for each class. Specifically, we first define a conventional Categorical Cross Entropy(CCE) for training a DL model with total of n samples with c classes in a multi-class classification problem as:

$$CCE\ (\mathbf{x}, \mathbf{y}; \theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\delta_j(\mathbf{y}_i)\log f\ (\mathbf{x}_i; \theta) \qquad (8)$$

where $\mathbf{x_i}$ denotes the i-th training sample, $\mathbf{y_i}$ denotes a corresponding label, and θ denotes parameters of the DL model. The $f\ (\mathbf{x}_i; \theta)$ in (8) is a return from a soft-max function from feed-forward result of the network and the δ_j function creates

one-hot vector encoding for multi-class classification, which can be defined as follows:

$$\delta_j\ (\mathbf{y}_i) = \begin{bmatrix} 1 & if\ \ \mathbf{y}_i = j \\ 0 & otherwise \end{bmatrix} \tag{9}$$

The CCE Loss function in (8) produces a loss value by adding all the values of the log operation to the prediction probability of each class. The prediction probability of each class is determined by the $\delta_j(\mathbf{y}_i)$ function, which is determined by the label $\mathbf{y}_i$. That is, since the log-loss is accumulated in proportion to the frequency of the labels during the training process, it may negatively affect the result of the backpropagation algorithm in the imbalanced dataset. To compensate for such shortcomings of conventional categorical cross entropy, we propose an ACE loss as follows:

$$ACE\ (\mathbf{x}, \mathbf{y}, \theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} f(w_p^j w_d^j, \alpha)\delta_j(\mathbf{y}_i) log f\ (\mathbf{x}_i; \theta) \tag{10}$$

which contain two additional weight terms, i.e., $\mathbf{w_p^j}$ and $\mathbf{w_d^j}$, in addition to (8). We make these weights *adaptively* behave according to the precision and skewed class distributions in the emotion prediction. The weights are defined as

$$w_d^j = c\left(\frac{1 - n_j/n}{c-1}\right) \tag{11}$$

$$w_p^j = c\left(\frac{1 - h_j/n_j}{\sum_{j=1}^{c} 1 - h_j/n_j}\right) \tag{12}$$

$$f(w_p^j w_d^j, \alpha) = \frac{\sum_{j=1}^{c} w_p^j w_d^j}{c} + \alpha(w_p^j w_d^j - \frac{\sum_{j=1}^{c} w_p^j w_d^j}{c}) \tag{13}$$

where n_j is the number of samples of the j-th class, and h_j denotes the number of true positive (hit) samples of the j-th class in a training process. To improve the accuracy of minority classes caused by class imbalance, we adopt a ''precision compensation'' weight term w_p^j in (13), which is adversely proportional to precision for each class. Differently expressed but the intuition is similar to that of Focal Loss [34]: if the precision with respect to a minor class is low, the weight should become larger to minimize prediction error and balance the accuracy of the minor class. Conversely, when the precision with respect to a major class is high, the weight is decreased to avoid over-fitting towards a major class.

What differentiates our idea on ACE from the convention is the distribution compensation weight term in (11), i.e., w_d^j, which alleviates the asymmetry in the loss caused by the data label imbalance. It increases the loss value when the number of samples in the j-th class n_j is small, and vice versa when n_j is relatively larger than other classes. This is an important aspect for our CDA framework: since the numbers of samples in each dataset are *significantly different*, we address this issue by balancing the numbers with w_d^j. Moreover, because the number of samples for each emotion class is also skewed, we applied the same scheme to both and dataset classifier and emotion classifiers. The function $f(w_p^j w_d^j, \alpha)$ adjusts how much to concentrate on imbalanced data between CCE and ACE by α. (it becomes CCE when α is 0). We expect that these adaptive parameters will improve the performance of overall accuracy by balancing the precisions of each class.

In practice, when training a DL algorithm, a loss function typically is calculated for each mini-batch. If we define the two weight terms of the ACE loss function for each mini-batch, n is defined as the size of each mini-batch. Then, the weight value of the ACE loss function for each mini-batch is determined based on the statistical property of label distribution and class accuracy, which are adaptively changed for each mini-batch.

Combining the above two ideas, the final proposed method uses the ACE as the loss of the emotion classifier in the CDA structure. That is, L in the equations in (1) and (2) are replaced with (10) to complete our model.

III. EXPERIMENTAL RESULTS

Our experiments were performed based on three different in-the-wild datasets with facial expression images, which are the most popularly used in facial expression recognition. The description on the datasets as well as the experimental setting, results, and validations are given below.

A. DATASETS

We used various in-the-wild datasets with 7 different emotion classes. The 7 classes consist of 6 basic emotions, i.e., Happiness, Sadness, Disgust, Anger, Fear and Surprise, and Neutral. The three datasets that we used consists of facial expression images that were collected from the web by searching with emotion-related keywords.

RAF Dataset: The Real-world Affective Faces (RAF) dataset contains image data from Flicker (https://www.flickr.com) by parsing emotion keywords that are related to the 7 emotion classes. There was a total of 315 annotators who labeled the images based on their knowledge of psychology. Specifically, each image and its label were validated by 40 independent annotators to increase the credibility of the dataset. The dataset consists of 15, 339 annotated images where 12, 271 belongs to the training set and the remaining 3, 068 images are assigned to the testing set [22].

ExpW Dataset: Facial Expression in-the-Wild (ExpW) is a dataset collected by a search engine using emotion-related keywords. It consists of 88, 600 facial expression images and does not offer a train set and a test set separately [24]. Thus, we randomly separated 3, 500 images as a test set and remaining 85, 100 images as a train set.

AffectNet Dataset: AffectNet is a facial expression dataset collected by searching 1,250 emotion-related keywords from search engines such as Google, Bing, and Yahoo. Each image was assigned to one annotator and labeled. There are a total of 287, 401 images related to 7 classes, including a train set of 283, 901 images and a validation set of 3, 500 images [23]. A test set has not been made public yet, and thus we used this validation set as a test set in all our experiments.

Figure 3 shows the facial expression class distributions in training and testing sets from the three datasets. In the case

TABLE 1. Comparisons of performances with the state-of-the-art methods.

Method	Pretrain	Train Set	Backbone	Align	Aff Test	RAF. Test
Occ. Razor [35]	-	Aff.	ResNet-18	N	-	80
ECAN [36]	2.6M	RAF2.0	VGG-Face	N	51.84	-
IPA2LT [29]	-	RAF+Aff.	ResNet-80	N	57.31	86.77
gACNN [37]	ImageNet	RAF	VGG-16	Y	-	85.07
gACNN [37]	ImageNet	Aff.	VGG-16	Y	58.78	-
Sep. Loss [38]	ImageNet	RAF	ResNet-18	N		86.38
Sep. Loss [38]	ImageNet	Aff.	ResNet-18	N	58.89	
Cov. Pool. [39]	-	RAF	CNN	Y	-	**87.00**
Ours	-	RAF+Exp.+Aff.	ResNet-50	N	**61.57**	86.83

FIGURE 3. Class distribution of RAF, ExpW, and AffectNet datasets. Blue: training samples, Red: testing samples. Happiness and Neutral are dominant in all three training sets, but AffectNet dataset has a uniform testing set.

of RAF and ExpW, we see that class imbalance is reflected in both the training set and the testing set. However, in the case of AffectNet, while the class label distributions in the training set are unbalanced, the class labels in the testing set are distributed evenly. We will focus on the test set of AffectNet since we want a classifier that is not biased towards a few specific expressions.

B. EXPERIMENTAL DESIGN

We used 50-layer ResNet [40] as our backbone network in all experiments. Throughout our experiment, the batch-size was 2048, the number of train-epochs was set to 120, and the stochastic gradient descent optimizer was used as an optimization method with a GPU, NVIDIA's TITAN RTX. In order to evaluate dataset bias only and remove other nuisance factors, we *did not use* the aligned images provided from each dataset; we only used face region information provided from each dataset and cropped facial images from the original images. Then, all images were resized to the same size of 100×100 which were used as the inputs to our framework. For data augmentation, random-cropping and horizontal-flipping were applied to all images so that the quality of all images become randomized. Only for the experiment in section III-C, we used the batch-size to 350.

C. CLASSIFYING FACIAL EXPRESSIONS IN IMAGES

We first present the main result from a facial expression recognition experiment comparing performance of our result to those from state-of-the-art baselines. Table 1 shows the summary of the performances. We combined the three datasets (i.e., RAF, AffectNet and ExpW) to create a large-scale multi-site dataset and trained our model. In this experiment, we used ExpW dataset as a part of training dataset but excluded for evaluation and comparisons with other methods, because it does not provide a public testing set. As seen in Table 1, our framework showed the best performance on the test set from AffectNet, which returned 2∼9% improvement on the recognition accuracy. Considering that the distribution of emotion labels are balanced in the test set of AffectNet dataset but all the models being compared here are trained with skewed distribution of labels in the training set, the result shows that our model is able to properly learn and generalize even with substantial class imbalance. Moreover, notice that the accuracy with our model is even better than the stat-of-the-art methods (with pre-training) trained on AffectNet only. This is important since training a classical DL model with the multi-site data decrease the performance of the model, which is to be shown with a separate experiment (Table 2 in section III-E). Here, we still achieve good accuracy with the combined multi-site dataset by minimizing dataset biases.

TABLE 2. Accuracy from cross-dataset facial expression recognition experiment (Training/Testing on different datasets).

Train	RAF Test	Exp. Test	Aff. Test	Avg.
RAF	**82.07**	54.54	46.20	60.94
Exp.	73.92	**70.97**	43.09	62.66
Aff.	76.21	63.86	**58.89**	66.32
Avg.	77.40	63.12	49.39	-

(Exp.: ExpW, Aff.: AffectNet)

We obtained comparable accuracy to the state-of-the-art result on RAF test data [39] with only 0.17% difference. We think that the difference is coming from the conventional face alignment used in other state-of-the-art baselines, which is typically used to reduce large unwanted variations on in-the-wild datasets. We did not apply the face alignment in the preprocessing since it may behave as a separate covariate affecting the original biases in each dataset.

D. IDENTIFYING BIAS IN MULTI-SITE FACIAL EXPRESSION DATASET

In order to confirm that there exists clear bias between different datasets, we first performed 1) dataset classification and 2) cross-dataset recognition experiments to quantify the bias introduced from the three facial expression datasets. In the dataset classification experiments, we exclusively used 381, 272 images as a training set and 10, 068 images as a test set by combining the three datasets, i.e., RAF, ExpW, and AffectNet.

Since these datasets were collected from natural scenes (i.e., from the wild), we first hypothesized an ideal condition that there wouldn't be any dataset bias between these datasets which would yield randomly classified result (with 33% accuracy). We performed the dataset classification

TABLE 3. Architecture of label extractor, emotion classifier, and domain classifier. Note that the input dimension for the dataset classifier is 1, 007, which is the dimension of a concatenated vector of a image feature (1, 000) and a label (7).

Label Extractor	Emotion Classifier	Dataset Classifier
Input1 × 7	Input 1 × 1000	Input 1 × 1007
Conv1 × 1 × 512, DropOut, LeakyReLU	Conv1 × 1 × 512, LeakyReLU	Conv1 × 1 × 512, LeakyReLU
Conv1 × 1 × 512, DropOut, LeakyReLU	Conv1 × 1 × 256, LeakyReLU	Conv1 × 1 × 256, LeakyReLU
Conv1 × 1 × 512, DropOut, LeakyReLU	Conv1 × 1 × 7	Conv 1 × 1 × 3
Conv1 × 1 × 512, DropOut, LeakyReLU		
Conv1 × 1 × 256, DropOut, LeakyReLU		
Conv1 × 1 × 7		

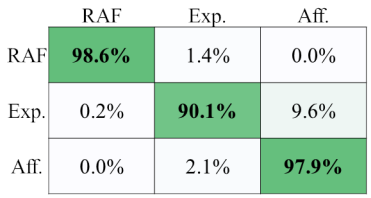


FIGURE 4. Confusion matrix of the dataset classification experiment. Surprisingly, the three in-the-wild datasets were distinguished with high accuracy demonstrating there exists dataset bias.

with ResNet-50, which is a part of our framework as a feature extractor. Unexpectedly, we obtained 91.74% dataset classification accuracy despite the dataset classification was performed on in-the-wild datasets. Even though the differences caused by image quality and alignment from each dataset were minimized, the unfortunate high classification performance was achieved clearly demonstrating the existence of dataset selection bias in each in-the-wild dataset. Especially, dataset bias of RAF dataset was the largest among the three datasets, followed by AffectNet and the ExpW as shown in a resultant confusion matrix in Figure 4. In case of RAF dataset, there were 40 annotators validating the images and expressions; perhaps there was a policy for determining expression labels that the annotators were made to fit which made the dataset bias stronger with repeated validation process.

In addition, we performed experiments on cross-dataset recognition (i.e., training/testing on different datasets) to validate generality of each dataset and baseline performance of the ResNet-50 algorithm trained on each dataset. Table 2 shows the results of facial expression classification with cross-dataset setting on the three in-the-wild datasets.

We observed a large performance drop when training and testing were performed on different datasets as shown in the off-diagonals of Table 2. Conversely, when training and testing processes were done on the same dataset, as shown in the diagonal elements of Table 2, we obtained much higher classification performances.

In particular, training with RAF dataset showed steepest drop when tested with test sets from other datasets with the lowest generalization capability compared to other datasets, whereas AffectNet dataset had the least performance drop. This was expected considering the size of training samples in descending order of AffectNet, ExpW, and RAF. Regarding the mean of classification accuracies as a measure classification difficulty for each dataset, classification with the test sets became more difficult in the order of AffectNet, ExpW, and RAF. Interestingly, we obtained the lowest classification performance with AffectNet; this may be because the distribution of images for each class in the training set and the test set are significantly different. In other words, since the class distributions in the train sets from the three datasets are skewed and the class distribution of the test set from AffectNet is balanced, this imbalanced distribution between a train set and the test set may negatively affect overall performances on test experiments with AffectNet dataset.

Throughout these two experiments, we confirmed that there exist clear biases among these three in-the-wild datasets, as well as other potential challenges with skewed class distributions. These are critical challenges for utilizing cross-datasets (i.e., multi-site dataset) scheme to setup a large-scale facial expression classification experiment to improve performance of DL algorithms; simply merging different datasets may hurt the performance of the DL framework. Our framework precisely tackles these issues to achieve successful results.

E. ANALYSIS ON RECOGNITION PERFORMANCE WITH MULTI-SITE DATASET

The purpose of this cross-dataset experiment is to validate improved generalization ability of our framework by running it over heterogeneous images from different datasets. We expect the performance to improve as we utilize more number of training samples by combining multiple datasets (i.e., cross-dataset) with the ideas we proposed in section II. Despite a larger dataset size, merging different datasets requires careful control of dataset bias that come from different dataset properties. In the following, we show experimental results with a multi-site dataset combining RAF, ExpW, and AffectNet datasets. For these experiments, we used the CDA architecture with a feature extractor using the ResNet-50 as a backbone, a label extractor, three emotion classifiers, and a dataset classifier as shown in Table 3. The λ_1 for the feature extractor and λ_2 for the label extractor were set to increase from 0 to 1 during the entire training epochs. The rest of the parameters were set as the same as in section III-B.

Different deep learning models were trained using all training samples from all three in-the-wild datasets. Three models are compared: 1) baseline method trained only on AffectNet, 2) baseline method trained on the three datasets, and 3) our proposed method that uses the CDA strucutre. As shown in Table 4, the proposed method achieved better performance than the others on all the test sets. Comparing the two baseline methods, although it showed

TABLE 4. Comparisons of facial expression recognition accuracy.

Train Set	Method	Model	RAF Te.	ExpW Te.	Aff. Te.	Acc. Avg.	F1. Avg.
Aff.	Baseline	ResNet-50	76.21	63.86	58.89	66.32	47.33
RAF+ExpW+Aff.	Baseline	ResNet-50	84.94	72.03	56.80	71.26	57.98
RAF+ExpW+Aff.	**Ours**	CDA	**86.60**	**73.89**	**61.03**	**73.83**	**59.13**

(Te.: test set, Aff.: AffectNet, Acc.: Accuracy, F1.: F1 score)

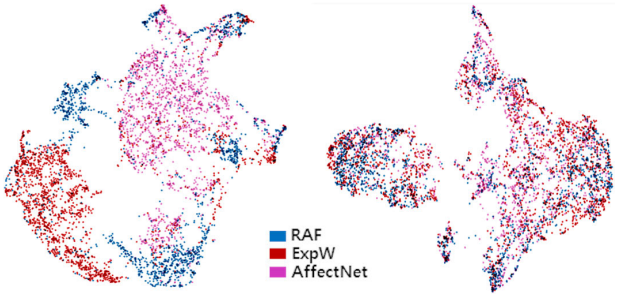


FIGURE 5. UMAP visualization of feature vectors. Left panel: using the baseline method, Right panel: using the proposed method. While the three datasets are clearly distinguished in the feature space (left) with dataset bias, our model is trained to combine them as a single dataset (right).

better average accuracy when trained on the entire three datasets than the case with training using AffectNet training set only, the individual performance on the AffectNet test set decreased. That is, despite the larger sample size by merging the three in-the-wild datasets, the selection and annotation bias of the datasets were well highlighted, resulting in poor performance rather than complementary effects on the AffectNet test set; we concluded that simply combining training sets from the multiple datasets does not produce meaningful results for generalization. On the other hand, using our proposed method, we gained increase in the accuracy for all the test sets despite the training sets were a mixture of different datasets.

In Figure 5, distribution of trained features from baseline with combined dataset (left) and the proposed method (right) are compared using Uniform Manifold Approximation and Projection (UMAP) [41] to show the effects from the proposed feature extractor. In the left of Figure 5, clear selection bias for each dataset is observed with clusters of distributions for each dataset in the trained features, while the features from our feature extractor show mixed distribution to make the three dataset as a single global dataset.

Figure 6 expands the results with AffectNet in more detail with recall matrices obtained from the three methods. Looking at the recall matrix in the middle column, true positive rate per class and recall values on the diagonals of the recall matrix are slightly decreased compared to the baseline method in the first column (Mean Recall Scores: Baseline with AffectNet 0.59 ± 0.20, Baseline with the combined dataset 0.57 ± 0.20). The diagonals representing recall is higher as the green becomes darker. Notice that the green colors in the last columns (i.e., elements in the last column) in each recall matrix in Figure 6 for Neutral are degrading across baseline with the combined datasets and our proposed method. This means the proposed method improves generalization ability using the training samples from all three datasets, where dataset bias is decreased compared to other methods. The proposed method was effectively applied to the AffectNet test set, with higher and more balanced true positive rates than the baselines. The mean recall score of the proposed method was 0.61 ± 0.15.

Notice that the precision per class using the proposed framework shown at the bottom of each matrix is much more balanced than other methods. This is highly desirable considering that the AffectNet dataset has uniform distribution of labels in the testing set. In particular, we even get balanced number of predictions for *neutral* class. We want to emphasize this result, since we hypothesized that *the dataset bias in neutral class may be the strongest* among the seven classes of facial expression. This may be because, when a policy to annotate or select facial expression images as neutral class is ambiguous, it would result in relatively small inter-class variation compared to other classes [42]. Using our proposed method, the number of neutral predictions was reduced from 981 in the baseline with the three datasets to 733. This means that the range of prediction for neutral class is reduced compared to those of the baseline algorithms, where more strict criteria was applied using the proposed method for predicting the neutral class. Although different datasets have different opinions on neutral facial expression, predictions based on our proposed method are *narrowed down to common characteristics* of the three independent datasets.

In addition, the precision for neutral class has improved despite the decrease in the number of predictions for neutral class using our framework. This is because the number of false positives in terms of neutral class has decreased significantly compared to the baseline methods. Such result suggests that the proposed method appropriately drove our predictions towards reducing dataset biases from the different datasets for neutral class. Overall performance of the proposed method was improved as seen in the mean of precisions (Baseline with the three datasets: 0.61 ± 0.11, Proposed method with the three datasets: 0.64 ± 0.11).

We also performed visualization of feature vectors using the UMAP method to see how well the extracted feature vectors are representing the seven emotions. In Figure 7, the left panel shows the result from the baseline method where the seven classes for facial expression are not clearly distinguished with mixed distribution; we observe samples only in the happiness class (brown tags) form a cluster. However, in the right panel showing our result, we can see that the emotions are separately distributed at intervals to distinguish those clusters. Notice that the Neutral samples are clustered at the center of the distribution, which is desirable since different emotions manifest from the Neutral status.

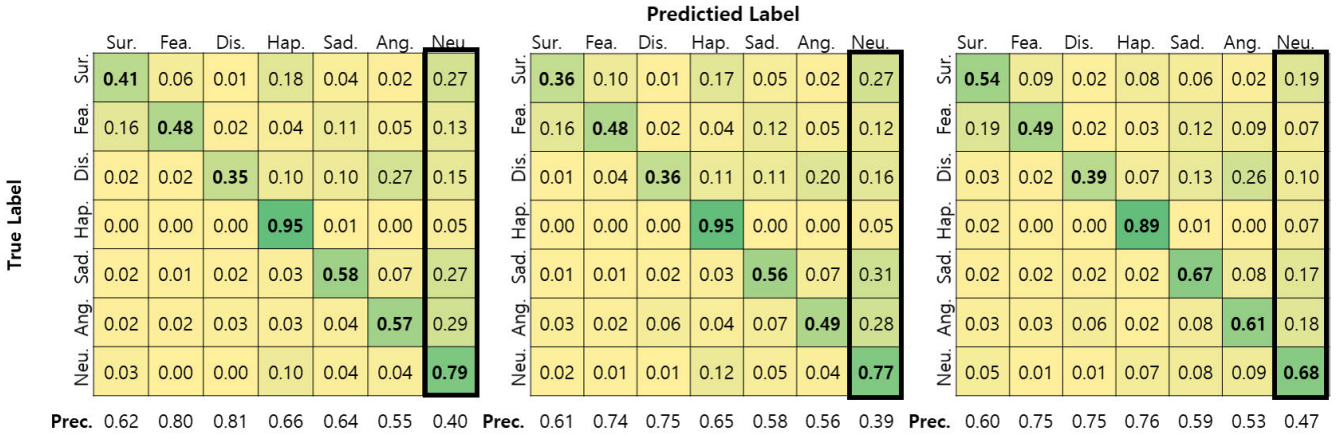


FIGURE 6. Recall matrices and Precision from expression recognition on the test set from AffectNet. Left: baseline trained on the AffectNet, Middle: baseline trained on the combined datasets, Right: proposed method on the three datasets. Notice that the prediction result with our method is much more evenly distributed than others on the AffectNet test set as shown in the diagonal elements of each recall matrix. The darker the color in the diagonals, the better true positive rate (recall).

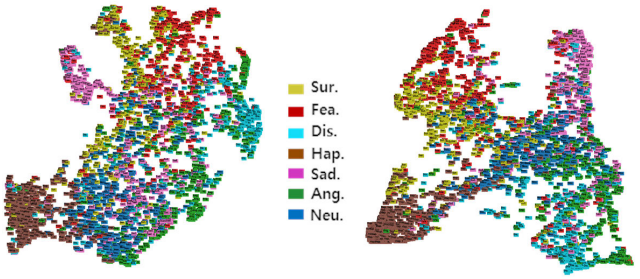


FIGURE 7. UMAP visualization distribution of samples in the feature vector space. Left panel: from the baseline method, Right panel: from the proposed method. The colors denote facial expressions Fear(red), Anger(green), Neutral(blue), Sadness(pink), Disgust(cyan), Surprise (dark yellow). Our result show better clustering of emotions with the Neutral (blue) being at the center of distribution.

IV. CONCLUSION

We designed a cross-dataset adaptation scheme for combining multiple datasets to retain sufficient sample size to train a DL model and minimize biases that exist across different datasets. Our model is designed to learn the "general" representation of data in specific classes that exist across multiple datasets. We applied our framework to a facial expression recognition problem using three independent in-the-wild datasets which had large dataset biases and class imbalance problem. We confirmed that this is a serious bias between datasets with dataset classification analyses and demonstrated extensive empirical results to evaluate the performance of generalization ability of our framework. We achieved improved performance over state-of-the-art algorithms with a balanced test set, i.e., AffectNet, when trained with multiple independent training sets with skewed class distribution as well as comparable results on RAF test dataset. There is a great potential that our method can be applied to various domains where combination of multi-site datasets is required to acquire enough data.

REFERENCES

[1] E. Redcay, D. Dodell-Feder, M. J. Pearrow, P. L. Mavros, M. Kleiner, J. D. E. Gabrieli, and R. Saxe, "Live face-to-face interaction during fMRI: A new tool for social cognitive neuroscience," *NeuroImage*, vol. 50, no. 4, pp. 1639–1647, May 2010.

[2] C. H. Liu and W. Chen, "Beauty is better pursued: Effects of attractiveness in multiple-face tracking," *Quart. J. Experim. Psychol.*, vol. 65, no. 3, pp. 553–564, Mar. 2012.

[3] A. Bulling and H. Gellersen, "Toward mobile eye-based human-computer interaction," *IEEE Pervas. Comput.*, vol. 9, no. 4, pp. 8–12, Oct./Dec. 2010.

[4] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 696–706, May 2002.

[5] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[6] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jan. 1991, pp. 586–587.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Brit. Mach. Vis. Conf. (BMVC)*, 2015, pp. 1–12.

[8] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1155–1164.

[9] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 394–405, Feb. 2011.

[10] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7832–7841.

[11] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016.

[12] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Web-scale training for face identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2746–2754.

[13] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 919–928.

[14] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2636–2640.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

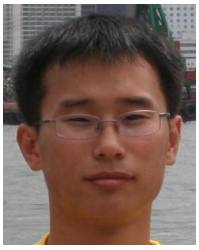
[16] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[17] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[18] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–8.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[20] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: Challenge of recognizing one million celebrities in the real world," *Electron. Imag.*, vol. 2016, no. 11, pp. 1–6, Feb. 2016.
[21] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, Jun. 2011, pp. 1521–1528.
[22] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2852–2861.
[23] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.
[24] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning social relation traits from face images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3631–3639.
[25] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 884–906, Jun. 2019.
[26] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1306–1315.
[27] T. Tommasi and T. Tuytelaars, "A testbed for cross-dataset analysis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014.
[28] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[29] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 222–237.
[30] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos, "Cross-database facial expression recognition based on fine-tuned deep convolutional network," in *Proc. 30th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2017, pp. 405–412.
[31] W. Zheng, Y. Zong, X. Zhou, and M. Xin, "Cross-domain color facial expression recognition using transductive transfer subspace learning," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 21–37, Jan. 2018.
[32] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1180–1189.
[33] S. Li and W. Deng, "Deep emotion transfer network for cross-database facial expression recognition," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3092–3099.
[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[35] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie, "An Occam's razor view on learning audiovisual emotion recognition with small training sets," in *Proc. Int. Conf. Multimodal Interact. (ICMI)*, 2018, pp. 589–593.
[36] S. Li and W. Deng, "A deeper look at facial expression dataset bias," 2019, *arXiv:1904.11150*. [Online]. Available: http://arxiv.org/abs/1904.11150
[37] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.
[38] Y. Li, Y. Lu, J. Li, and G. Lu, "Separate Loss for Basic and Compound Facial Expression Recognition in the Wild," in *Proc. Asian Conf. Mach. Learn. (ACML)*, 201, pp. 897–911.
[39] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 367–374.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 630–645.
[41] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*. [Online]. Available: https://arxiv.org/abs/1802.03426
[42] C. Bell, C. Bourke, H. Colhoun, F. Carter, C. Frampton, and R. Porter, "The misclassification of facial expressions in generalised social phobia," *J. Anxiety Disorders*, vol. 25, no. 2, pp. 278–283, Mar. 2011.

BYUNGOK HAN received the B.S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2008, and the M.S. degree in robotics and the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2010 and 2016, respectively. Since October 2016, he has been with the Electronics and Telecommunications Research Institute (ETRI), South Korea, as a Senior Researcher. His research interests include computer vision, machine learning, pattern recognition, and human–robot interaction.



WOO-HAN YUN received the B.S. degree in electronic and electrical engineering from Sungkyunkwan University, Suwon, South Korea, in 2004, and the M.S. degree in computer science and engineering from POSTECH, Pohang, South Korea, in 2006. He is currently pursuing the Ph.D. degree with KAIST, Daejeon, South Korea. He joined the Electronics and Telecommunications Research Institute (ETRI), in 2006. His current research interests include object detection, recognition, and human–robot interaction.

JANG-HEE YOO (Senior Member, IEEE) received the B.Sc. degree in physics and the M.Sc. degree in computer science from the Hankuk University of Foreign Studies, South Korea, in 1988 and 1990, respectively, and the Ph.D. degree in electronics and computer science from the University of Southampton, U.K., in 2004. He was a Visiting Scientist with the University of Washington, Seattle, USA, from August 2014 to July 2015. Since November 1989, he has been with the Electronics and Telecommunications Research Institute (ETRI), South Korea, as a Principal Researcher. He has also been a Professor with the Department of Computer Software, University of Science and Technology, South Korea. His current research interests include computer vision, human motion analysis, biometric systems, HCI, and intelligent robot. He is a member of IEEK.

WON HWA KIM (Member, IEEE) received the B.S. degree in information and communication engineering from Sungkyunkwan University, in 2008, the M.S. degree in robotics from KAIST, in 2010, and the Ph.D. degree in computer sciences from the University of Wisconsin–Madison, in 2017. He has been a Young Professional of IEEE, since 2019. He is currently an Assistant Professor in computer science and engineering with The University of Texas at Arlington (UTA), Texas, USA. Prior to joining UTA, he worked as a Researcher with the Data Science Team, NEC Laboratories America, in 2017. His work has been published in top-tier AI conferences, such as NeurIPS, CVPR, ICCV, ECCV, MICCAI, ISBI, as well as in high impact journals, such as *NeuroImage*, *NeuroImage: Clinical*, *Brain Connectivity*, and *Brain Imaging and Behavior*. His research interests include computer vision, machine learning and neuroscience, and developing novel methods for analyses of unconventional data.