Face2Exp: Combating Data Biases for Facial Expression Recognition

Dan Zeng[1], Zhiyuan Lin[1], Xiao Yan[1], Yuting Liu[2], Fei Wang[3], Bo Tang*[1]
[1] Research Institue of Trustworthy Autonomous Systems, Southern University of Science and Technology,
Department of Computer Science and Engineering, Southern University of Science and Technology
[2] JD.com, Beijing, China; [3] School of Microelectronics, Southern University of Science and Technology
{zengd@,12132456@mail.,yanx@,wangf@,tangb3@}sustech.edu.cn;lalenaliu17@gmail.com

Abstract

Facial expression recognition (FER) is challenging due to the class imbalance caused by data collection. Existing studies tackle the data bias problem using only labeled facial expression dataset. Orthogonal to existing FER methods, we propose to utilize large unlabeled face recognition (FR) datasets to enhance FER. However, this raises another data bias problem—the distribution mismatch between FR and FER data. To combat the mismatch, we propose the Meta-Face2Exp framework, which consists of a base network and an adaptation network. The base network learns prior expression knowledge on class-balanced FER data while the adaptation network is trained to fit the pseudo labels of FR data generated by the base model. To combat the mismatch between FR and FER data, Meta-Face2Exp uses a circuit feedback mechanism, which improves the base network with the feedback from the adaptation network. Experiments show that our Meta-Face2Exp achieves comparable accuracy to state-of-the-art FER methods with 10% of the labeled FER data utilized by the baselines. We also demonstrate that the circuit feedback mechanism successfully eliminates data bias [1].

1. Introduction

Facial expression recognition (FER) has many applications in human-computer interaction and affective computing [1, 2, 7]. However, as shown in Figure 1 (a), existing FER training datasets are biased towards some majority classes, which leads to poor test accuracy for the minority classes. The bias is because some facial expressions (e.g., contempt, disgust) are rare in daily life and it is expensive collect many samples for them. Deep neural networks (DNNs) trained with biased data tend to favor majority classes and perform poorly on minority classes.

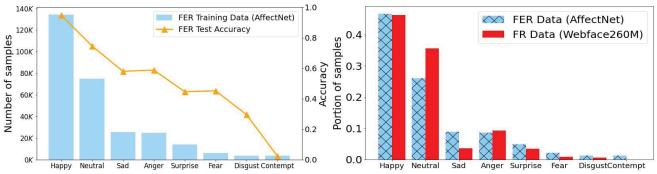


Figure 1. (a) The class distribution of FER training data is biased, resulting in different test accuracy among the classes. (b) The class distributions of FER and FR data mismatches. We use ResNet50 trained with the entire AffectNet to test FER and FR data.

Early FER methods, such as [9] and [19], train deep neural network with biased FER data to classify expressions and observe sharp accuracy drop for minority expression categories. To tackle the class bias, later methods [32], [10] use human facial movements (i.e., facial action unit, facial landmarks) as side information for expression recognition. However, these methods require datasets with labels besides expression, which are even more expensive to collect. Some FER methods [26, 29] improve performance by removing ambiguous samples from the training set.

To summarize, existing FER methods work on biased FER datasets with focuses such as model design, side information and difficult samples. However, it is well-known that high quality training data is crucial to the performance of DNN models. Methods [6, 21, 24] enhance FER with small-scale controlled unlabeled data to improve performance. As large-scale class-balanced FER dataset is expensive to acquire, we propose to use large face recognition (FR) datasets without expression labels to improve FER. For example, Webface260M [42], MS-Celeb-1M [13] and VGGFace2 [3] are all million-level FR datasets, which contain face images with good comprehensive variety (i.e., varied pose, identities, varying illumination, and different expressions). In contrast, the largest public FER dataset contains only 440K images. However, this inevitably raises another data bias problem because FER data and FR data have mismatched distributions as illustrated in Figure 1(b).

To combat the forgoing data biases, we propose the

[1]Code is available at link:https://github.com/danzeng1990/Face2Exp.
* Corresponding author.

Meta-Face2Exp framework, which utilizes unlabeled **face** data to enhance **exp**ression recognition through the **meta** optimization framework. The meta-learning objective is to minimize the loss function of the model predictions on the challenging facial expressions, conditioned on the balanced FER data. Meta-Face2Exp consists of two networks, namely *a base network* and *an adaptation network*, which are connected via a circuit feedback paradigm that uses the feedback from an adaptation network to improve a base network for de-biased knowledge extraction. We refer to each full-trained base and adaptation model as a generation. At each generation, the adaptation network is trained on large-scale FR data using pseudo labels generated by the base network in order to distill rich facial expression knowledge (in meta-train phase). The base network learns prior facial expression knowledge on de-biased FER data, which are sampled to ensure class balance. To tackle the mismatch between FR and FER data, our circuit feedback paradigm informs the base network how good the pseudo labels are by using the cognitive difference of the adaptation network on biased FR data and de-biased FER data in meta-test phase. If their cognitive difference is large, the base network is punished to use the adverse direction of current gradients. Thus, the base network is continuously improved, more convincing pseudo expression labels are generated for training the adaptation network, and finally the adaptation network learns de-biased expression knowledge.

To sum up, this paper makes the following contributions:

- We explain two data biases, i.e., class imbalance in FER data and class distribution mismatch between FR and FER data, which inspired Meta-Face2Exp, the first work to utilize large-scale unlabeled FR data to enhance FER. We think Meta-Face2Exp provides a general framework to utilize large-scale unlabeled FR data for other face related tasks (e.g., gender/race classification, age estimation) that lack high quality data.
- We propose the Meta-Face2Exp framework to extract de-biased knowledge from auxiliary FR data through the meta optimization framework. Meta-Face2Exp provides an cost-effective paradigm for facial expression recognition.
- We conduct extensive experiments on widely-used FER benchmarks including AffectNet [22] and RAF-DB [19] to demonstrates the effectiveness of our Meta-Face2Exp framework. Specifically, Meta-Face2Exp obtains comparable results to state-of-the-art FER methods using the only 10% of labeled FER data.

2. Related Work

In this section, we discuss recent work related to FER and learning with unlabeled data. We then highlight our key idea of exploring large-scale unlabeled FR data to enhance FER which is different from existing methods.

2.1. Facial Expression Recognition

Methods for single-dataset FER can be divided into three categories, i.e., deep classification network based methods, human facial movements based methods, and ambiguous expression annotation based methods.

Deep classification network based methods: They adopt deep neural network trained with labeled FER data to preserve local similarity and maximize inter-class scatters for discriminative feature extraction [9, 10, 19, 36]. FaceNet2ExpNet [9] designs a two-stage training algorithm. In the pre-training stage, the face recognition network is used to train the convolutional layer, and then the expression labels are executed to fully train the network. However, their results suffer from a sharp drop in the minority expression categories. DLP-CNN [19] uses a locality preserving loss to pull together the locally neighboring faces of the same class, and a softmax loss to force different classes to be separated. Very recently, DACL [10] has proposed a new loss to adaptively learn discriminative features. TransFER [36] has proposed a new architecture based on Transformer to learn relation-aware local representation.

Human facial movements based methods: They provide strong clue of expression movements to help learn discriminative expression features and attention mechanism is usually explored [10, 30, 32]. RMT-Net [4] establishes the connection between FER and facial landmark localization through association learning and residual learning, so that training data with single-task labels can be used for the multi-task network. FDRL [25] consists of a feature decomposition network (FDN) and a feature reconstruction network (FRN). FDN decomposes the basic features into a set of facial action-aware latent features based on the facial action unit. FRN reconstructs the expression feature by learning intra-feature relation weight and inter-feature relation weight for each latent feature. However, these methods usually require the dataset with multi-task label.

Ambiguous expression annotation based methods: They typically tackle with intractable labeling confusion to enhance FER result, which are most close to our work. IPA2LT [40] is the first work to learn an FER model with multiple inconsistent annotated data and large-scale unlabeled data. LDL [5] converts one-hot facial expression labels into label distribution to solve the problem of annotation inconsistency and learns these distributions from auxiliary tasks including action unit recognition and facial landmark detection. SCN [29] adopts ranking regularization to weight each training sample and relabel these lowest ranked group to suppress the uncertainties of FER. DMUE [26] introduces an auxiliary multi-branch learning for latent distri-

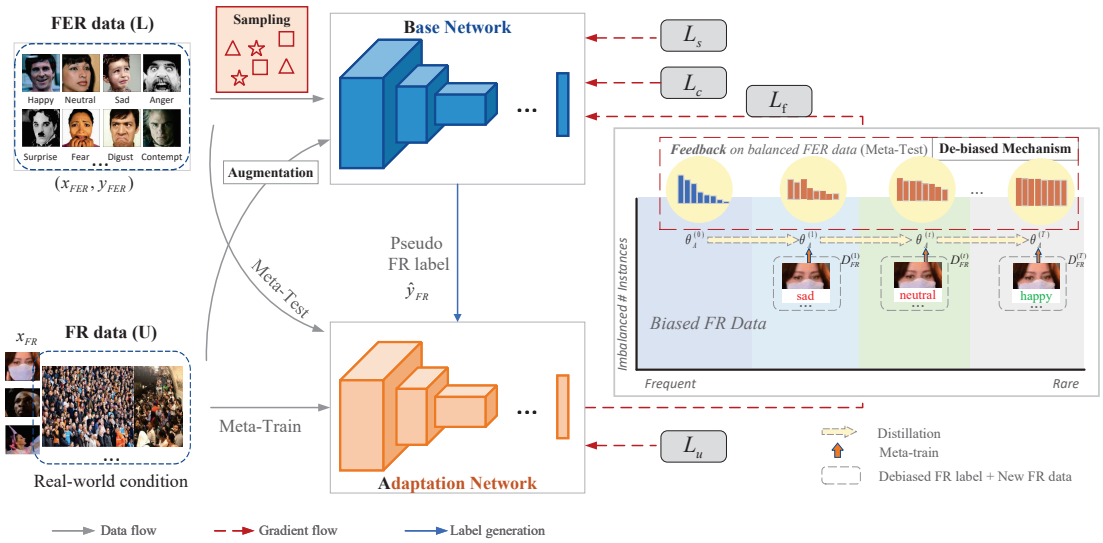


Figure 2. An illustration of the proposed Meta-Face2Exp framework. Meta-Face2Exp consists of an adaptation network and a base network. In the meta-train phase, the adaptation network is trained on unlabeled FR data with pseudo FR label. In the meta-test phase, the adaptation network estimates the cognitive differences between biased FR data and de-biased FER data to update the base network as the feedback. Two networks are connected in a circuit feedback paradigm to extract de-biased knowledge from large-scale FR data.

bution mining and uses pairwise features to estimate uncertainty in order to resolve the ambiguous nature of expression annotations.

2.2. Learning with Unlabeled Data

In cross-dataset FER, semi-supervised learning has been extensively studied for learning with unlabeled data. Existing works [11, 12, 37, 38] transfer the knowledge from the training set to the target data to jointly learn the optimal nonlinear discriminative features on both source and target datasets. In contrast, Ref. [31] uses the target data generated by GAN to fine-tune the model trained on the source data to minimize the difference between source and target expressions. AdaFER [28] uses objective facial action units to perform auxiliary training of unsupervised domain adaptation to relieve the annotation bias between source and target domains. AGRA [35] combines graph representation propagation and adversarial learning to fine-grained adaption to local features according to data inconsistency and bias between different datasets. ECAN [18] learns domain-invariant and discriminative feature representations by applying maximum mean discrepancy as re-weighted regularization and class-conditional regularization learning. *Compared to cross-data FER, our work focuses on exploring large-scale unlabeled FR data (i.e., from different application) to enhance FER which is more challenging.*

Some general methods such as pseudo label [17], noisy student [34] are proposed to learn with unlabeled data. Fixmatch [27] simplifies the learning process by training the model with high-confidence pseudo labels. UDA [33] improves semi-supervised learning by incorporating data augmentation [8] to limit the invariance of model predictions to input noise. Recently, meta learning methods have been used to purify pseudo labels. MPL [23] enables the teacher network to adjust based on student's performance feedback on labeled data. CPGML [39] proposes inexactly-supervised meta-learning. The training samples have only coarse-grained labels to reduce the need for data annotation. *In Meta-Face2Exp, we use meta-learning philosophy to address data biases including class imbalance and distribution mismatch to enhance FER. Specifically, it can be regarded as a bilevel optimization problem [15], which is primarily concerned with networks learned at different levels: an inner- and an outer-level optimization. Base network, the outer level optimization, learns knowledge to perform well on FER validation sets after training. Adaptation network, the inner level optimization, provides the feedback to improve base network according to its cognitive differences between biased FR data and de-biased FER data.*

3. Meta-Face2Exp

3.1. Framework Overview

As illustrated in Figure 2, the proposed Meta-Face2Exp framework for facial expression recognition consists of an adaptation network ($\mathcal{A}$) and a base network ($\mathcal{B}$). These two networks have the same network architecture with independent weights and are connected with a circuit feedback paradigm. At each generation, the adaptation network uses generated pseudo labels $\hat{y}_{FR}$ (i.e., by utilizing the base network) on unlabeled data x_{FR} for training. The base network learns a prior expression knowledge on the labeled

data x_{FER} which are sampled to ensure class balance by the sampling module $\texttt{Smp}(\cdot)$. With the de-biased mechanism of Meta-Face2Exp, the base network is gradually improved based on the feedback of the adaptation network according to the cognitive differences between biased FR data and de-biased FER data. As a result, the base network can produce better pseudo labels for training the adaptation network in the next generation. For example, as illustrated in the right part of Figure 2, we estimate the initial adaptation network (i.e., trained with long-tail FER data) and observe a severely skewed blue accuracy distribution. Later, we can observe more and more flattened accuracy distribution (i.e., red accuracy distributions) from training step 1 to T with de-biased mechanism. Meanwhile, the predicted FR label is generally corrected from sad, neutral, to happy expression based on the feedback on balanced FER data. By design, the two networks constantly complement each other to extract de-biased knowledge in the FER task. During training, the $\mathcal{A}$ network and the $\mathcal{B}$ network are updated alternatively. In the inference stage, only the adaptation model $\mathcal{A}$ is used for facial expression prediction.

3.2. Adaptation Network ($\mathcal{A}$)

For the adaptation network, large-scale unlabeled FR data is exploited to enhance FER because FR data has abundant and comprehensive variety. As illustrated in Figure 2, Meta-Face2Exp trains the adaptation network by encouraging two networks to predict similar conditional classification distribution on unlabeled FR data with loss $\mathcal{L}_u$:

$$\mathcal{L}_u = \text{CE}(\hat{y}_{FR}, \mathcal{A}(x_{FR}; \theta_{\mathcal{A}})). \tag{1}$$

To minimize the cross-entropy loss on pseudo FR label $\hat{y}_{FR}$ (i.e., a one-hot target label) which is the expression with the highest score derived from $\mathcal{B}(x_{FR}; \theta_{\mathcal{B}})$. Unlike ground-truth labels, pseudo labels change dynamically during training. The parameters of the adaptation network $\theta_{\mathcal{A}}$ are updated in the meta-train phase. In the meta-test phase, the balanced FER dataset (i.e., for training the base network) is used to estimate the cognitive differences between biased FR data and de-biased FER data.

3.3. Base Network ($\mathcal{B}$)

For the base network, labeled FER images x_{FER} is used to train the network and y_{FER} is used as the ground truth label. We first adopt a sampling module $\texttt{Smp}(\cdot)$ to ensure the class distribution of FER data are balanced. Specifically, we randomly select the same number of samples on each facial expression class, which ensures balanced classes are generated for training base network. As illustrated in Figure 2, three losses of supervised loss, consistency loss, and feedback loss are proposed to guide the learning process of base network which can be expressed as

$$\mathcal{L}_{\mathcal{B}} = \mathcal{L}_s + \mathcal{L}_c + \mathcal{L}_f. \tag{2}$$

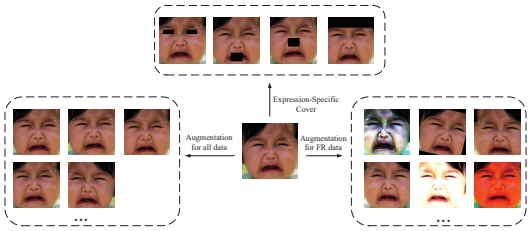


Figure 3. An illustration of augmented facial images.

Specifically, supervised loss and consistency loss only work with base network while feedback loss considers the performance of meta-test on adaptation network.

Supervised learning with FER data: With loss $\mathcal{L}_s$, Meta-Face2Exp trains the base network to minimize the cross-entropy loss on labeled yet balanced FER data:

$$\mathcal{L}_s = \text{CE}(y_{FER}, \mathcal{B}(x_{FER}; \theta_{\mathcal{B}})), \tag{3}$$

where $\theta_{\mathcal{B}}$ are the parameters for the base network of Meta-Face2Exp and CE represents the cross-entropy loss.

Consistency learning with FR data: In addition, we apply augmentation module $\texttt{Aug}(\cdot)$ on large-scale FR data to pay more attention to expression-sensitive face regions. Meta-Face2Exp trains the base network to guarantee a consistent conditional distribution between the original FR data and the augmented data by utilizing loss $\mathcal{L}_c$:

$$\mathcal{L}_c = \text{CE}(\mathcal{B}(x_{FR}; \theta_{\mathcal{B}}), \mathcal{B}(\texttt{Aug}(x_{FR}); \theta_{\mathcal{B}})). \tag{4}$$

We propose an effective yet expression-specific augmentation method $\texttt{Aug}(\cdot)$ for FR data. These augmented images are not only used for training adaptation network but also used for consistency learning in base network. As for consistency learning, base network requires original images and augmented counterparts to have close class conditional distribution. There are three types of augmentations for image generation as illustrated in Figure 3 including conventional transformations (randomly crop, rotation, and horizontal flip) for all data on the left box, extensive image transformation (i.e., rotation, erasing, and pixel-wise image processing) for FR data on the right box and expression-specific augmentation on the top box. Considering facial expression is closely related to facial landmarks, we augment face images to purify facial expression feature extraction by covering areas unrelated to facial expression. Specifically, we apply MTCNN [41] to detect five facial landmarks and empirically determine the patch centered on landmarks, i.e., 50×20 pixels for eyes, 50×40 pixels for nose and mouth, and 224×50 pixels for the forehead.

Feedback learning with FR data: In Meta-Face2Exp, the base network $\mathcal{B}$ and adaptation network $\mathcal{A}$ are updated through the circuit feedback paradigm. Specifically, $\mathcal{B} \rightarrow \mathcal{A}$ is linked with pseudo label generation and $\mathcal{A} \rightarrow \mathcal{B}$

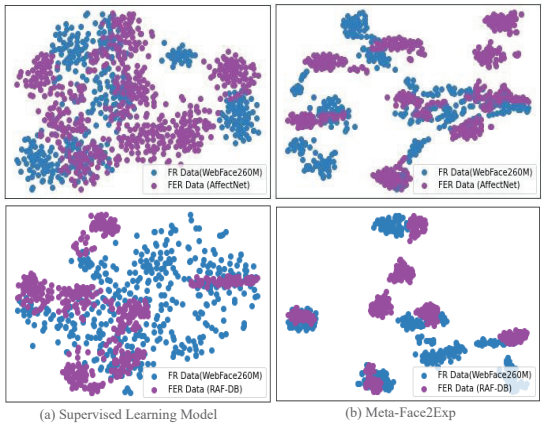
is linked with feedback loss $\mathcal{L}_f$. The feedback loss worked on the base network can be expressed as

$$\mathcal{L}_f = f \cdot \mathrm{CE}(\hat{y}_{FR}, \mathcal{B}(x_{FR}; \theta_{\mathcal{B}})), \tag{5}$$

where f estimates the feedback of cognitive difference between FR and FER data to help update the parameters of the base network. The definition of feedback coefficient f can be expressed as

$$f = \eta_{\mathcal{A}} \cdot (\nabla_{\theta_{\mathcal{A}}^{(t+1)}} \mathrm{CE}(y_{FER}, \mathcal{A}(x_{FER}; \theta_{\mathcal{A}}^{(t+1)}))^{\top} \cdot \nabla_{\theta_{\mathcal{A}}} \mathrm{CE}(\hat{y}_{FR}, \mathcal{A}(x_{FR}; \theta_{\mathcal{A}}^{(t)}))), \tag{6}$$

where f is expressed as a dot product of two terms. The first term: the gradients of the *new* adaptation network on de-biased FER data. The second term: the gradients of the *old* adaptation network on biased FR data. If two terms have the same/different gradient sign, the base network is updated according to the same/adverse of the current gradients. The absolute value of the dot product determines the strength of the gradients updates. The adaptation network uses pseudo-labeled data to update the parameters to $\theta_{\mathcal{A}}^{(t+1)}$. In particular, we approximate it with the parameters obtained from $\theta_{\mathcal{A}}^{(t)}$ by updating the base network parameters on $(x_{FR}, \hat{y}_{FR})$, i.e., $\theta_{\mathcal{A}}^{(t+1)} = \theta_{\mathcal{A}}^{(t)} - \eta_{\mathcal{A}} \nabla_{\theta_{\mathcal{A}}} \mathrm{CE}(\hat{y}_{FR}, \mathcal{A}(x_{FR}; \theta_{\mathcal{A}}))$.



(a) Supervised Learning Model (b) Meta-Face2Exp

Figure 4. An illustration of De-biased mechanism combating distribution mismatch bias. FR data and FER data aligns better by using Meta-Face2Exp (b) than SL model (a).

De-biased Mechanism: Exploring auxiliary FR data to enhance FER will inevitably cause mismatched distribution between the FR and FER data. To verify the de-biased behavior of Meta-Face2Exp, we analyze the distribution of facial expression features between FR and FER data by tSNE. For FR data, we randomly generate total 415 samples from Webface260M which is biased distributed. For FER data, we generate 560 samples from AffectNet and RAF-DB testing set with every expression category 70 facial images, respectively. Figure 4 presents the typical results of our experiment, where purple and blue colors represent FER and FR data, respectively. The supervised learning (SL) model is trained with all AffectNet. The SL model shows a severe mismatch between de-biased FER and biased FR data in Figure 4(a). Specifically, the blue color in Webface260M and purple color in AffectNet and RAF-DB span the entire feature space and there is no obvious cluster of two colors that are well aligned, which motivates our de-biased design.

We show the philosophy of how the de-biased mechanism works. The behavior of new adaptation network on balanced FER data is used as the measurement. Specifically, when adaptation network on FR data and new adaptation network on FER data have the same cognitive (i.e., sign of their gradients), we obtain positive feedback coefficients (i.e., positive value for $\mathcal{L}_f$), which encourages the update of base network by using the current direction of gradients. When adaptation network on FR data and new adaptation network on FER data have different cognitive (i.e., sign of their gradients), feedback coefficients will give a negative sign, which punishes the update of base network by using the adverse direction of current gradients. In this way, feedback is used as the reward signal that goes back through the base network, determining how parameters of the base network impact the gradient of the adaptation network for de-biased feature extraction. As a result, the learning of adaptation network on biased FR data will perform de-biased behavior that is consistent with the evaluation procedure on balanced FER data. As Figure 4(b) illustrated, we observe the distribution mismatch has been greatly alleviated with de-biased mechanism and the feature layout of de-biased FR data (i.e., Webface260M) is similar to that in balanced FER data. For example, two distributions overlap at the center bottom, indicating de-biased knowledge is learned even in biased FER data. Because FER data contains basic expressions and FR data has compound expressions as illustrated in Figure 6. This explains their features cannot be perfectly aligned even with our method.

3.4. Algorithm for Meta-Face2Exp

We listed detailed step-by-step pseudo-code for Meta-Face2Exp in Algorithm 1. Meta-Face2Exp extracts de-biased knowledge from auxiliary FR data via a circuit feedback paradigm. At each generation, the adaptation network is firstly updated by minimizing the unsupervised loss $\mathcal{L}_u$ illustrated in line 7. As a result, $\mathcal{B} \rightarrow \mathcal{A}$ is linked with pseudo label generation for FR data. The base network is consequently updated by utilizing three losses (i.e., supervised loss $\mathcal{L}_s$, consistency loss $\mathcal{L}_c$, and feedback loss $\mathcal{L}_f$) illustrated in line 17. Specifically, three losses for guiding the learning process of base network are illustrated in line 9, line 11, and line 15, respectively. As a result, $\mathcal{A} \rightarrow \mathcal{B}$ is linked with the feedback loss from line 12 to 15. By de-

Algorithm 1 Training procedure of Meta-Face2Exp

Input: Labeled data $\mathcal{D}^{'}_{FER}$ and unlabeled data $\mathcal{D}_{FR}$
Outputs: $\Theta^{(T)}_{\mathcal{A}}$
Initialize: $\theta^{(0)}_{\mathcal{B}}$ and $\theta^{(0)}_{\mathcal{A}}$

Get balanced labeled data: $\mathcal{D}_{FER} \leftarrow \texttt{Smp}(\mathcal{D}^{'}_{\texttt{FER}})$
for $t = 0...T-1$ **do**
  $x_{FER}, y_{FER} \leftarrow$ SampleMiniBatch$(\mathcal{D}_{FER})$
  $x_{FR} \leftarrow$ SampleMiniBatch$(\mathcal{D}_{FR})$
  $\hat{y}_{FR} \leftarrow$ Forward$(x_{FR}, \theta^{(t)}_{\mathcal{B}})$
  Update the adaptation network using pseudo label:
  $\theta^{(t+1)}_{\mathcal{A}} \leftarrow \theta^{(t)}_{\mathcal{A}} - \eta_{\mathcal{A}} \nabla_{\theta_{\mathcal{A}}} \text{CE}(\hat{y}_{FR}, \mathcal{A}(x_{FR}; \theta_{\mathcal{A}}))$
  Compute the base network's gradient on FER data:
  $g^{(t)}_{\mathcal{B},s} \leftarrow \nabla_{\theta_{\mathcal{B}}} \text{CE}(y_{FER}, \mathcal{B}(x_{FER}; \theta_{\mathcal{B}}))$
  Compute the base network's gradient on FR data:
  $g^{(t)}_{\mathcal{B},c} \leftarrow \nabla_{\theta_{\mathcal{B}}} \text{CE}(\mathcal{B}(x_{FR}; \theta_{\mathcal{B}}), \mathcal{B}(\texttt{Aug}(x_{FR}); \theta_{\mathcal{B}}))$
  Compute the base network's feedback coefficients:
    applying Equation (6)
  Compute the base network's gradient via feedback:
  $g^{(t)}_{\mathcal{B},f} \leftarrow f \cdot \nabla_{\theta_{\mathcal{B}}} \text{CE}(\hat{y}_{FR}, \mathcal{B}(x_{FR}; \theta_{\mathcal{B}}))$
  Update the base network:
  $\theta^{(t+1)}_{\mathcal{B}} \leftarrow \theta^{(t)}_{\mathcal{B}} - \eta_{\mathcal{B}} \cdot (g^{(t)}_{\mathcal{B},s} + g^{(t)}_{\mathcal{B},f} + g^{(t)}_{\mathcal{B},c})$
end for
return $\Theta^{(T)}_{\mathcal{A}}$

sign, the two networks constantly complement each other to extract de-biased knowledge in the FER task.

To circumvent data biases, the use of circuit feedback is the key. First, the base network learns a prior expression knowledge from class balanced FER data, which results in more de-biased expression prediction during pseudo label generation ($\mathcal{B} \rightarrow \mathcal{A}$). Second, the adaptation network compares the cognitive differences (i.e., before and after updating parameters) on de-biased FER data to update the learning of base network by utilizing the feedback loss, which explicitly tackles with the class distribution mismatch between FR and FER data ($\mathcal{A} \rightarrow \mathcal{B}$). In the end, the adaptation network is equipped with the de-biased expression knowledge even without labeling.

4. Experiments

In this section, we first introduce the experiment settings and compare Meta-Face2Exp with existing state-of-the-art. We then show our circuit feedback successfully eliminates data bias. We finally conduct ablation study on loss design.

4.1. Datasets and Metrics

We use AffectNet [22] and RAF-DB [19] as our target FER dataset for facial expression recognition and choose Webface260M [42], the newest one as our FR data which are illustrated in Table 1. **AffectNet** is by far the most challenging and largest FER dataset, providing expression categories annotations. By querying expression-related keywords from three search engines, there are $440,000$ images collected from the Internet. Among them, the $280,000$ training images and $4,000$ testing images are manually annotated with eight facial expressions (e.g., neutral, happy, anger, sad, fear, surprise, disgust, and contempt). It has an unbalanced training dataset and balanced test dataset. **RAF-DB** is another large-scale FER dataset that contains $30,000$ facial images with seven basic or compound annotations (i.e., neutral, happy, surprise, sad, anger, disgust, and fear) annotated by about 40 independent taggers. **Webface260M** [42], a million-scale dataset, is by far the largest public FR dataset which contains noisy 260M faces from 4M identities and 42M clean faces from 2M identities. The *mean class accuracy* as well as *confusion matrix* are used for measurement. In addition, we also report the *standard deviation (std)* among the accuracy of each expression class to measure the FER bias.

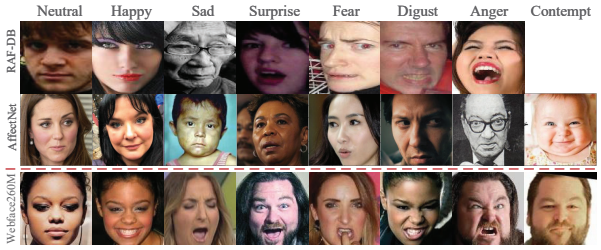


Figure 5. Example expressions on FER data (i.e., RAF-DB, AffectNet) and FR data (i.e., Webface260M). For FR data, manually select eight facial expressions of three identities for display.

Table 1. Dataset for deep face recognition and deep facial expression recognition.

Dataset	# Identities	# Exps	# Images	Publications
Webface260M (FR data)	4M	-	260M	CVPR'21
RAF-DB (FER data)	-	7	30K	CVPR'17
AffectNet (FER data)	-	8	440K	TAC'17

4.2. Implementation Details

Training Details: For AffectNet, we sample $28,608$ images as labeled FER data for training (only 10% of AffectNet) and $4,000$ images for testing. For RAF-DB, we use all $12,270$ images with seven basic expression (i.e., without sampling) for training and $3,068$ images are used for testing. We do not apply sampling because the minority expression only contains 281 images which are too small to train the network. All training face images are detected and resized to 256×256 pixels, and augmented by random cropping to 224×224 pixels. By default, we use ResNet50 [14] as the backbone for both base network and adaptation network. The learning rate is first initialized (i.e., 1e-2 for the

base network, 1e-3 for the adaptation network) and further decayed with cosine annealing strategy. Once the training is completed, we finetune the adaptation network with labeled dataset by a fixed learning rate of 1e-5. The batch size is set to 32. The entire training steps for training AffectNet and RAF-DB are $180,000$ and $30,000$, respectively. It is trained end-to-end with one Nvidia RTX2080 GPU.

Baselines: We compare Meta-Face2Exp with state-of-the-art baseline. We include 8 models that train with the entire labeled training dataset, i.e., SL [14], gaCNN [20], IPA2LT [40], RAN [30], CAKE [16], SCN [29], LDL [5] and DMUE [26]. Among them, 4 models (i.e., SL, IPA2LT, DMUE, SCN) provide both results on AffectNet and RAF-DB. IPA2LT is a pioneer work to solve the problem of inconsistent annotations. We also include the CAKE model that uses 7 classes for training and testing on AffectNet. All results are from their papers. For SL models, we train them under our experiment settings (i.e., ResNet50, 100% labeled data size) and conduct extensive hyper-parameter tuning for performance optimization.

4.3. Comparison with State-of-the-Art Methods

We report mean class accuracy of Meta-Face2Exp and the baseline models for comparison. The results in Table 2 show that Meta-Face2Exp outperforms state-of-the-art (i.e., CAKE) if model is trained and tested with 7 classes on AffectNet. For 8 facial expression classes, Meta-Face2Exp can still obtain comparable results (i.e., 60.17%) to state-of-the-art methods using the only 10% labeled FER data. According to results in Table 3, we are setting a new second best records on RAF-DB with mean accuracy of 88.54% and which is also comparable to state-of-the-art methods (i.e., 88.76%). We believe that an ideal FER system should report not only high mean accuracy but also low std accuracy. *Unfortunately, std accuracy is largely ignored by existing methods.* We will focus on analyzing and discussing the std accuracy for FER and show that Meta-Face2Exp can achieve high mean accuracy with much lower std accuracy.

Table 2. Comparison on AffectNet. $^+$ denotes both AffectNet and RAF-DB are used as the training data. $\star$ denotes the method is trained and tested with 7 classes. Only 10% labeled data is used in our method.

Methods	IPA2LT$^+$	RAN	CAKE($\star$)	DMUE	SCN	SL	Ours($\star$)
Accuracy(%)	55.71	59.50	61.7	<u>63.11</u>	60.23	58.37	**64.23**

Table 3. Comparison on RAF-DB. $^+$ denotes both AffectNet and RAF-DB are used as the labeled training data.

Methods	gaCNN	IPA2LT$^+$	LDL$^+$	DMUE	SCN	SL	Ours
Accuracy(%)	85.07	86.77	85.53	**88.76**	87.03	84.16	<u>88.54</u>

Table 4. Mean and Std accuracy on AffectNet for different models. All models are trained and tested with 7 expression categories.

Models	SL	Meta-Face2Exp(Ours)		
Labeled Data Size	100%	1%	5%	10%
Mean Accuracy(%)↑	58.37	53.54	<u>61.66</u>	**64.23**
Std Accuracy(%)↓	21.53	14.41	<u>10.69</u>	**10.07**

Table 5. Mean and Std accuracy on RAF-DB for different models.

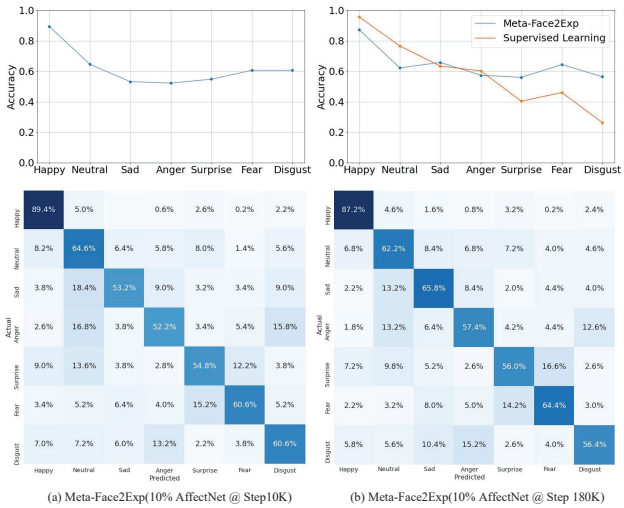
Models	SL	Meta-Face2Exp(Ours)		
Labeled Data Size	100%	25%	50%	100%
Mean Accuracy(%)↑	84.16	80.87	<u>85.04</u>	**88.54**
Std Accuracy(%)↓	15.48	**9.43**	10.70	<u>10.00</u>

4.4. Evaluation on class imbalance

Size of the labeled set: To verify the effectiveness of our Meta-Face2Exp for class imbalance, we report std accuracy of different models on AffectNet and RAF-DB in Table 4 and Table 5. Specifically, we use Meta-Face2Exp model accuracy as a function of the size of labeled dataset. For AffectNet, we use Meta-Face2Exp model accuracy as a function of 1%, 5%, and 10% *balanced* data. For RAF-DB, we do not require balanced data as it is a small-scale dataset and use 25%, 50%, and 100% of *unbalanced* labeled data. As results shown, (1) Meta-Face2Exp significantly reduces the std accuracy and consistently outperforms the baselines by a large margin which demonstrate the de-biased behavior of Meta-Face2Exp. Specifically, we reduce the std accuracy from 21.53% to 14.41% on AffectNet by using only 1% of balanced labeled AffectNet. We also reduce the std accuracy from 15.48% to 9.43% by using only 25% unbalanced RAF-DB. (2) We can achieve better mean accuracy than baselines by training with small size of labeled data (i.e., 5% labeled AffectNet, and 50% labeled RAF-DB). (3) We found that larger labeled data size leads to better mean accuracy, but not necessarily leads to lower std accuracy. It verifies that the circuit feedback can successfully eliminate data bias and is friendly to the very limited data (i.e., has nothing to do with the size of labeled training data).

We also compare the std accuracy of Meta-Face2Exp to other existing method. As SCN is the only open-source project, we test the std accuracy on RAF-DB by using the public available model. SCN achieves std accuracy of 15.56%, which is close to SL model with 15.48% but is around 6% higher than proposed Meta-Face2Exp in Table 5. This comparison again supports the effectiveness of Meta-Face2Exp to combat class imbalance.

De-biased behavior during training: We take Meta-Face2Exp model trained with 10% labeled AffectNet as an example to show the de-biased behavior during training procedure. The test accuracy of different training steps as well as confusion matrix between true and predicted labels are



(a) Meta-Face2Exp(10% AffectNet @ Step10K) (b) Meta-Face2Exp(10% AffectNet @ Step 180K)

Figure 6. An illustration of the de-biased behavior of Meta-Face2Exp. Blue curves from (a) to (b) show accuracy improvement of Meta-Face2Exp by training more steps. (b) Meta-Face2Exp obtains more balance yet higher accuracy compared to supervised learning model trained with 100% labeled data.

illustrated in Figure 6. First, we observe that the accuracy curve of different expressions gradually flatten with training as illustrated from Figure 6(a) to (b). Second, compare to SL model, proposed Meta-Face2Exp can greatly alleviate class imbalance as illustrated in Figure 6(b). For example, Meta-Face2Exp can achieve 56.4% on disgust expression while SL can only produce less than 30% of mean accuracy. Such de-biased behavior is benefit from our circuit feedback paradigm which continuously improves the mean accuracy of FER and learns de-biased facial expression knowledge.

Visualization Analysis: To further investigate effectiveness of our Meta-Face2Exp on FER and FR data, we show the prediction results of the supervised learning (SL) model and Meta-Face2Exp on eight expressions of AffectNet. Figure 7 illustrates some example faces and prediction results including predicted expression and probability from different facial expressions. Compared with the baseline trained with entire AffectNet, Meta-Face2Exp shows better recognition results. Specifically, our model predicts a higher probability of correctly recognizing facial expressions. From the middle two rows, our model can recognize facial expressions but the SL model cannot. From the bottom two rows, our model predicts a lower probability for facial expressions cannot be recognized.

4.5. Ablation Study

We explore the effects of Meta-Face2Exp using different loss functions. The ablation study of performance on AffectNet and RAF-DB is illustrated in Table 6. As results shown, (1) With four losses together, Meta-Face2Exp can achieve the best mean accuracy. (2) Unsupervised loss $\mathcal{L}_u$ enhances FER to a large margin, which verifies the effectiveness of pseudo label generation. (3) Consistency loss $\mathcal{L}_c$ plays a crucial role to enhance FER. This not only verifies the effectiveness of the proposed augmentation module (i.e., expression-specific cover), but also provides a clue that we can learn comprehensive knowledge from FR data. (4) Feedback loss $\mathcal{L}_f$ can further contribute to accuracy improvement which shows that de-biased mechanism works.

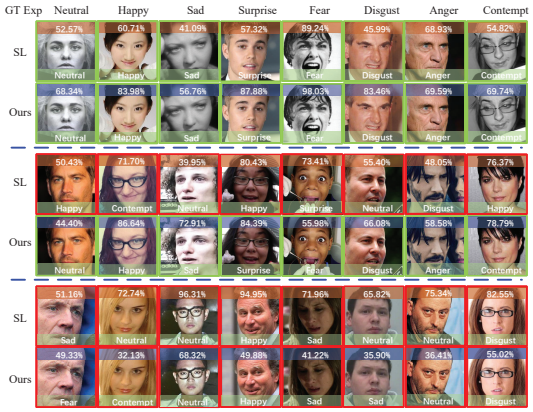


Figure 7. Comparison of Supervised Learning (SL) and our model.

Table 6. Performance of Meta-Face2Exp on 7 expression categories of AffectNet and RAF-DB with different loss functions.

Models	$\mathcal{L}_s$	$\mathcal{L}_u$	$\mathcal{L}_c$	$\mathcal{L}_f$	AffectNet	RAF-DB
1	✓	✗	✗	✗	60.66	84.16
2	✓	✓	✗	✗	62.03	86.25
3	✓	✓	✓	✗	63.60	87.26
4	✓	✓	✓	✓	**64.23**	**88.54**

5. Conclusions

In this paper, we propose Meta-Face2Exp for facial expression recognition, which utilizes unlabeled FR data to enhance FER through a meta optimization framework. It is inspired by the observations that FER data is class imbalanced and FR and FER data have a mismatched distribution. The key component is that the base network and the adaptation network constantly complement each other to extract de-biased knowledge through the circuit feedback paradigm. In particular, the de-biased mechanism can effectively produce low std and high mean accuracy. Experiments demonstrate that Meta-Face2Exp can obtain comparable results to state-of-the-art methods using the only 10% labeled FER data.

Acknowledgements

This work was supported by the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001) and the National Natural Science Foundation of China (No. 62176170, 62066042).

References

[1] Faiza Abdat, Choubeila Maaoui, and Alain Pruski. Human-computer interaction using emotion recognition from facial expression. In *2011 UKSim 5th European Symposium on Computer Modeling and Simulation*, pages 196–201. IEEE, 2011.

[2] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. Real time face detection and facial expression recognition: development and applications to human computer interaction. In *2003 Conference on Computer Vision and Pattern Recognition workshop*, volume 5, pages 53–53. IEEE, 2003.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[4] Boyu Chen, Wenlong Guan, Peixia Li, Naoki Ikeda, Kosuke Hirasawa, and Huchuan Lu. Residual multi-task learning for facial landmark localization and expression recognition. *Pattern Recognition*, 115:107893, 2021.

[5] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13984–13993, 2020.

[6] Ira Cohen, Fabio Gagliardi Cozman, Nicu Sebe, Marcelo Cesar Cirelo, and Thomas S Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553–1566, 2004.

[7] Jeff F Cohn and Fernando De la Torre. Automated face analysis for affective computing. 2015.

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[9] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.

[10] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2402–2411, 2021.

[11] Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–17. Springer, 2020.

[12] Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *BMVC*, page 104, 2019.

[13] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 87–102. Springer, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] Mike Huisman, Jan N Van Rijn, and Aske Plaat. A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541, 2021.

[16] Corentin Kervadec, Valentin Vielzeuf, Stéphane Pateux, Alexis Lechervy, and Frédéric Jurie. Cake: Compact and accurate k-dimensional representation of emotion. *arXiv preprint arXiv:1807.11215*, 2018.

[17] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[18] Shan Li and Weihong Deng. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, 2020.

[19] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2852–2861, 2017.

[20] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[21] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Enhancing expression recognition in the wild with unlabeled reference data. In *Asian Conference on Computer Vision*, pages 577–588. Springer, 2012.

[22] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[23] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021.

[24] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision (ECCV)*, pages 808–822. Springer, 2012.

[25] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7660–7669, 2021.

[26] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression

recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2021.

[27] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[28] Kai Wang, Yuxin Gu, Xiaojiang Peng, Panpan Zhang, Baigui Sun, and Hao Li. Au-guided unsupervised domain adaptive facial expression recognition. *arXiv preprint arXiv:2012.10078*, 2020.

[29] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020.

[30] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[31] Xiaoqing Wang, Xiangjun Wang, and Yubo Ni. Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Computational intelligence and neuroscience*, 2018, 2018.

[32] Zhengning Wang, Fanwei Zeng, Shuaicheng Liu, and Bing Zeng. Oaenet: Oriented attention ensemble for accurate facial expression recognition. *Pattern Recognition*, 112:107694, 2021.

[33] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[34] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[35] Yuan Xie, Tianshui Chen, Tao Pu, Hefeng Wu, and Liang Lin. Adversarial graph representation adaptation for cross-domain facial expression recognition. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1255–1264, 2020.

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

[37] Haibin Yan. Transfer subspace learning for cross-dataset facial expression recognition. *Neurocomputing*, 208:165–173, 2016.

[38] Keyu Yan, Wenming Zheng, Tong Zhang, Yuan Zong, Chuangao Tang, Cheng Lu, and Zhen Cui. Cross-domain facial expression recognition based on transductive deep transfer learning. *IEEE Access*, 7:108906–108915, 2019.

[39] Jinhai Yang, Hua Yang, and Lin Chen. Coarse-to-fine pseudo-labeling guided meta-learning for few-shot classification. *arXiv preprint arXiv:2007.05675*, 2020.

[40] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018.

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[42] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.