# Point Adversarial Self-Mining: A Simple Method for Facial Expression Recognition

Ping Liu, Yuewei Lin, Zibo Meng, Lu Lu, Weihong Deng, *Member, IEEE*, Joey Tianyi Zhou, and Yi Yang, *Senior Member, IEEE*

***Abstract*—In this article, we propose a simple yet effective approach, called point adversarial self mining (PASM), to improve the recognition accuracy in facial expression recognition (FER). Unlike previous works focusing on designing specific architectures or loss functions to solve this problem, PASM boosts the network capability by simulating human learning processes: providing updated learning materials and guidance from more capable teachers. Specifically, to generate new learning materials, PASM leverages a point adversarial attack method and a trained teacher network to locate the most informative position related to the target task, generating harder learning samples to refine the network. The searched position is highly adaptive since it considers both the statistical information of each sample and the teacher network capability. Other than being provided new learning materials, the student network also receives guidance from the teacher network. After the student network finishes training, the student network changes its role and acts as a teacher, generating new learning materials and providing stronger guidance to train a better student network. The adaptive learning materials generation and teacher/student update can be conducted more than one time, improving the network capability iteratively. Extensive experimental results validate the efficacy of our method over the existing state of the arts for FER.**

***Index Terms*—Facial expression recognition (FER), in-the-wild data, point adversarial attack.**

## I. Introduction

FACIAL expression analysis aims to comprehend the underlying human emotions and establish efficient communications between humans and humans or humans and computers [1]–[4]. Due to its emerging applications in human–computer interaction, facial expression recognition (FER) has received massive interest among the research community. In the past decade, the FER accuracy has been boosted significantly with the rapid development of modern convolutional neural networks (CNNs) [5].

As a “data-hungry” method, CNNs usually require a huge amount of annotated data for parameter learning. Existing FER datasets can be categorized into two groups: 1) lab-controlled and 2) in-the-wild datasets. In lab-controlled datasets, such as CK+ [6], the collecting environment is highly controlled, for example, frontal exaggerated expressive faces with limited occlusions and minimal illumination changes. These lab-controlled datasets have been widely adopted for evaluating proposed methods [7]–[17]. The limited sample number and small variations in those lab-controlled datasets make it difficult to train a deep network with high generalities. To improve model generalization abilities, researchers collect data in a more challenging environment (in-the-wild), and annotate those in-the-wild data with facial expression labels. Compared to their lab-controlled counterparts, samples from in-the-wild datasets, which contain spontaneous head poses and various occlusions, can better reflect data distribution in the real world. Therefore, in-the-wild datasets for FER are attracting more research interests in the research community.

To improve performance of FER on those in-the-wild datasets, various CNNs-based methods have been proposed [18]–[20]. Generally, those works manually design advanced architectures by utilizing an attention mechanism to simulate the human perception system. The attention map can be generated either by following a weakly supervised object localization strategy [18] or by the occlusion score generated from an occlusion detector [19]. For example, Zhang *et al.* [18] designed a weakly supervised local-global relation network to generate attention maps, indicating the facial regions crucial to the prediction. Li *et al.* [19] detected face landmarks and utilized the landmark confidence scores as the corresponding occlusion level. Based on the confidence score, the network can explicitly place appropriate focus on different facial regions to make a more accurate prediction. All those previous works need to introduce additional modules, increasing model sizes and inference complexity.

From a different point of view, we consider to boost the FER accuracy by simulating the human learning process. Unlike previous works modifying model architectures, we improve the

Manuscript received January 22, 2021; revised May 8, 2021; accepted May 28, 2021. This work was supported in part by the Agency for Science, Technology and Research (A*STAR) through AI and ANALYTICS SEED GRANT under Grant Z20F3RE005, and in part by the AME Programmatic Funding Scheme under Project A18A1b0045. This article was recommended by Associate Editor L. Jiao. *(Corresponding author: Joey Tianyi Zhou.)*

Ping Liu and Joey Tianyi Zhou are with the Institute of High Performance Computing, Agency for Science, Technology, and Research, Singapore (e-mail: joey.tianyi.zhou@gmail.com).

Yuewei Lin is with the Department of Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11973 USA.

Zibo Meng is with the U.S. Research Center, InnoPeak Technology Inc., Palo Alto, CA 94303 USA.

Lu Lu is with the Key Laboratory of Medical Molecular Virology, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China.

Weihong Deng is with the Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China.

Yi Yang is with the Centre for Artificial Intelligence, University of Technology Sydney, Sydney, NSW 2007, Australia.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3085744.

Digital Object Identifier 10.1109/TCYB.2021.3085744

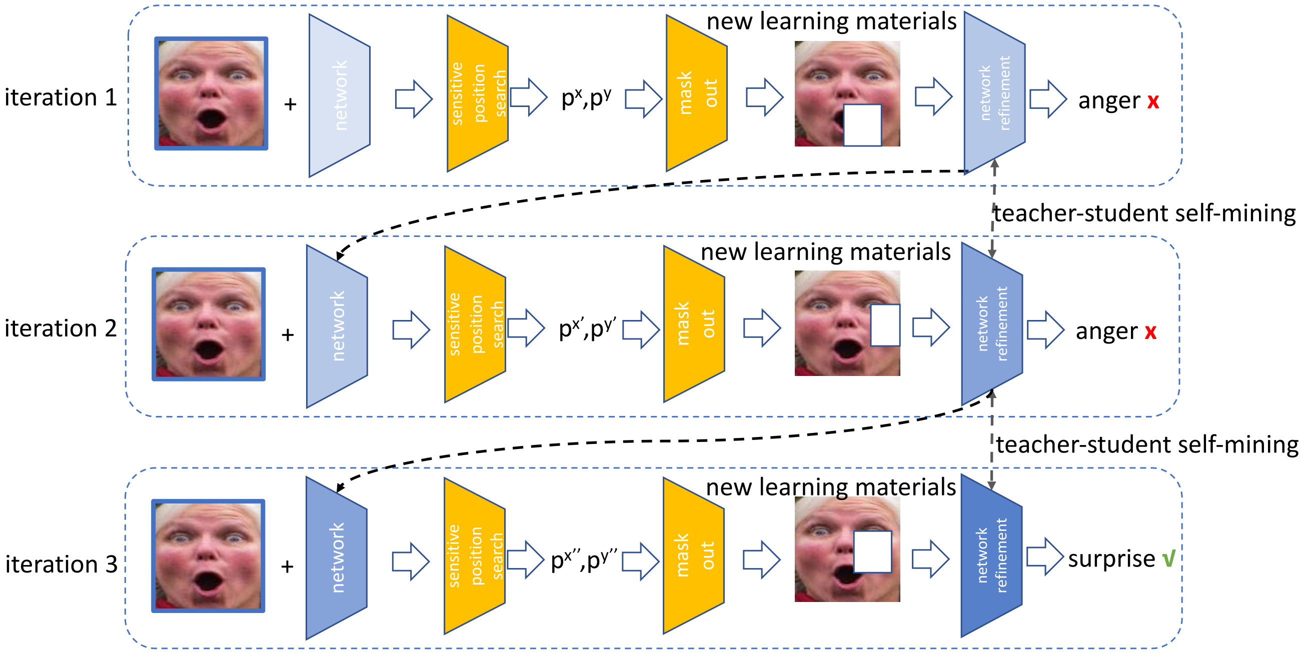


Fig. 1. Pipeline of the proposed PASM. To improve network recognition rate on FER, PASM keeps generating new learning materials for the network and providing guidance from an updated teacher network, which was a student in previous iterations. By taking advantage of new learning materials and updated teachers, PASM can progressively boost the recognition accuracy on facial expression datasets. Best viewed in color.

model capability by iteratively providing new learning materials and more advanced teachers. We argue that the same or even better performance can be achieved, if we appropriately utilize the training samples in hand. After all, the key information related to the target task already exists in the given samples.

To this end, we propose a self-mining framework to boost the recognition accuracy for FER and name it point adversarial self-mining (PASM). The proposed PASM integrates the idea of the point adversarial attack and teacher–student optimization strategy, making them collaborate together in an iterative way. In each iteration, the point adversarial attack module locates the most informative regions in each given sample. Because each sample has its own statistical distribution, the most sensitive region obtained by point adversarial attack is highly adaptive and has a direct influence on the classification. For each image, the located region is controlled by two factors: 1) the statistical distribution of the image itself and 2) the network for locating the region. By erasing the located region in the corresponding sample, a new image, which is treated as a "harder" learning material, is generated. Given those updated learning materials, the network has to spend more efforts to mine more knowledge to improve. In the following rounds/epochs, the improved network becomes an advanced teacher, providing guidance to train a better student network. Our PASM has three main steps: 1) train a teacher network on given samples; 2) locate the most informative region for each training sample by the teacher network trained in step 1), and then erase the located region to update learning samples; and 3) train a student network on those new learning materials and under the guidance from the teacher network. The steps 2) and 3) can be conducted more than one time, by treating the student obtained in the previous iteration as a teacher in the current iteration. When the learning process finishes, the final student network is with high generalization ability, which is demonstrated in our experimental results. The entire iterative process is illustrated in Fig. 1.

Our proposed PSAM is inspired by two previous related works: 1) random erasing [21] and 2) adversarial erasing [22], where subregions were erased from the original "clean" image to produce new samples to refine networks. However, there are a few significant differences between our method and [21] and [22]: 1) Reference [21] randomly chooses the subregions to erase, without considering the characteristics of the input and the network prediction capability. In other words, it is possible that the selected region has little discriminative information, and erasing them will not have any influence on the final prediction. On the contrary, our method behaves more reasonably since it locates the most informative region by considering the statistical information of each sample and the trained network capability; 2) unlike [21] working as a pure data augmentation strategy, our method integrates harder sample generation and the teacher–student optimization strategy into a unified self-mining framework. The two parts in our self-mining framework are conducted iteratively and work collaboratively; and 3) compared to [22] using an attention map to select regions, the region selected by our method is more structured and sparse, which aligns with the previous finding [15], [23]: in face analysis problems, not all facial regions but only a few of them contribute to the final predictions.

To sum up, our main contributions in this work are as follows.

1) We propose a simple yet effective method, called PASM, to improve the recognition accuracy of FER. Compared to previous works [24]–[26] designing specific architectures, our method does not bring any additional parameters or computation cost in inference stages, while still achieves higher or comparable performance.
2) Our method simulates the studying process in human societies. The method progressively generates new learning materials and provides guidance from a teacher network keeping updated. In PASM, the new learning materials are generated in an adaptive manner, considering both the statistical information of original learning materials and the teacher network status. Benefiting from those updated learning materials and teachers, the learning capability of the student network keeps improving.
3) We conduct extensive experiments on challenging facial expression datasets collected under real-world settings. Our method, although simple, achieves better or comparable performance compared with the previous methods utilizing complex architectures or dedicated loss functions.

## II. Related Work

In this section, we will elaborate on previous works that are the most related to our work, including FER approaches, data-augmentation strategies, and adversarial attack methods.

### *A. Facial Expression Recognition*

FER is an image-level classification research topic, which has been considered as a combination of three major steps: 1) feature learning; 2) feature selection; and 3) classifier construction [15]. Before the dominance of CNNs in this field, hand-crafted feature-based methods have been exhaustively studied. As elaborated by the recent survey papers [27]–[29], various hand-crafted features, such as Gabor-wavelet-based features [8], [9], [11]; histograms of oriented gradients (HOGs) [12], [13]; and local binary pattern (LBP) [30]–[32], have been developed and well demonstrated for data collected under lab-controlled settings. In the past decade, as CNNs have shown promise in different computer vision tasks, researchers in FER start to shift their attention from hand-crafted features to deep features [15], [16], [19], [33]–[35]. Comparing to their hand-crafted contemporaries, whose designing heavily depends on human expertise, deep features can be learned in a data-driven manner and have better performance on challenging datasets. To further improve the performance of CNNs for FER, various architectures have been developed and studied, such as deep belief networks [15], ResNets [16], InceptionNets [33], and generative adversarial networks (GANs) [34], [36], [37]. Most recently, researchers found that utilizing information from other modalities, such as brain waves [38] and audio [39], [40], can also boost the emotion recognition performance. For a systematic review for deep learning in FER, refer to [41].

### *B. Data Augmentation*

Data augmentation is one of the strategies that can effectively prevent deep networks from overfitting. As the architectures of the deep networks become deeper and more complicated, the number of parameters has increased dramatically. Without enough labeled data, it is easy for those networks to overfit and lose the generality. In particular, as pointed out in [21], in an extreme case, an overfitted model might achieve perfect accuracy for the training data while performing poorly on unseen data. To deal with the overfitting problem, various data augmentation strategies have been proposed and employed [21]. The basic idea of data augmentation is to introduce more variations into the training data without changing the statistical distribution of the original data. The most common and frequently used data augmentation techniques include random cropping, random flipping, and random color jittering. The efficacy of those three data augmentation strategies have been demonstrated in [42] and [43]. In the past two years, two novel data augmentation methods, namely, mixup [44] and random erasing [21], have been proposed. Mixup [44] combines pairs of examples and corresponding labels in a convex manner in order to generate new training samples. Random erasing [21] randomly selects a position in a given input and erases pixels around the selected position. All of those proposed data augmentation methods are complementary and can be combined together to train a deep neural network, which has been experimentally proved [45].

There are significant differences between PASM and random erasing [21]: 1) random erasing [21] selects the position to erase in a purely random manner, while our method selects the erasing position by considering the data statistical information and network capabilities; 2) reference [21] improves the model capability in a data augmentation manner, while our method works by simulating the human learning process: providing new learning materials and advanced teachers, both of which are updated in an adaptive manner; and 3) our method works in a progressive way, continuously improving the model capability.

### *C. Adversarial Attack*

With the successful application of deep neural networks in various computer vision tasks [46]–[51], CNNs have been deployed in more safety-critical scenarios [52], [53], such as autonomous driving, financial fraud detection, etc. However, recent works [54] pointed out that deep neural networks are not as stable as originally expected. On the contrary, they are vulnerable to adversarial examples, which are intentionally designed to mislead a trained deep neural network to make incorrect predictions. In previous works, adversarial examples are synthesized based on the model capability and given input [55].

Generally, adversarial attack can be categorized into two groups: 1) white-box attack [56]–[58] and 2) black-box attack [59], [60]. The difference between them is the availability of network information, for example, parameters, gradients, etc. Specifically, in a white-box attack setting, those network information are available to the attackers; while in a black-box

**Algorithm 1:** PASM

```
Input: input image set $\{\mathbf{x}_i, y_i\}$, $1 \leq i \leq N$; image sizes
       $W_i$ and $H_i$; patch size $S$; iteration number $iter$
Output: network parameter $w^*$
initialize a teacher network:
training target network with input images $\{\mathbf{x}_i, y_i\}$, get an
  initialization $w$ for a teacher network;
for $ind = 1, ..., iter$ do
    for $i$ = 1, ..., $N$ do
        generating new learning materials:
        sample an image $\{\mathbf{x}_i, y_i\}$ from the given set;
        for the sampled $\{\mathbf{x}_i, y_i\}$, utilize Equation. (3) to
          generate the sensitive position in it;
        mask out the region centered at the searched
          location;
        set the pixel value in the masked out region by
          the perturbed RGB value, generating a new
          learning sample $\mathbf{x}_i^{new}$;
        refine a student network and update the
        teacher network:
        leverage the new learning sample $\{\mathbf{x}_i^{new}, y_i\}$ and
          the teacher network, refine a student network
          based on (4);
    set the student network as a teacher network for the
      next iteration.
return $w_{ind}$ as $w^*$;
```

attack, it is unavailable to adversaries. Correspondingly, in a white-box attack, the model architecture and parameters are utilized to create the adversarial samples, which can attack the model to the most, while in the black-box attack, a feed-and-query strategy is the first choice for adversaries. The most frequently used white-box attack policies include the fast gradient sign method (FGSM) [56], Deep Fool [57], projected gradient descent (PGD) [58], etc. Recently, researchers have conducted a few works toward the black-box attack. Jiang *et al.* [60] proposed to extend natural evolution strategies to estimate the gradient for the black-box image attack. Papernot *et al.* [61] studied the feasibility of utilizing the adversarial example transferability for the black-box attack on static images. Readers can refer to [55] for a systematic review of the adversarial attack.

## III. Methodology

This section illustrates the details of the proposed PASM for FER. As illustrated in Algorithm 1, at the first step, we train a network based on given training samples. The trained network is utilized to generate updated learning samples in the second step, and acts as a teacher network in the third step. At the second step, the most sensitive position in each image is located by a point adversarial attack method, and then, the local regions around the located position are masked out to generate an updated sample. At the third step, given the guidance from the teacher network, a student network is trained based on the updated learning materials generated in the second step. Steps 2 and 3 can be repeated more than one time, in which the student network obtained the previous epoch/iteration acts as an advanced teacher in the current epoch/iteration.

### A. Initialize Teacher Network

In the first step, we train a network based on the original training samples, that is, $\{x_i, y_i\}$, where $x_i$ denotes a sample and $y_i$ denotes the corresponding label. The target network is denoted as $F(w)$, where $w$ denotes the network parameters. In order to conduct a fair comparison, we choose ResNet-34 [62] and VGG16 [43] pretrained by ImageNet [63] as an initial teacher in our experiment. In our training, we reset the output number to fit our interesting class number, that is, 7. We choose a cross-entropy (CE) loss in this step, which is formulated as

$$\min_{w} \frac{1}{N} \sum_{i}^{N} \mathrm{CE}((F(w, x_i), y_i)) \tag{1}$$

where $N$ denotes the number of training samples, and $w$ denotes the network parameters.

### B. Sensitive Location Search via Point Adversarial Attack

Given the initial teacher network, we locate the most sensitive position in each training sample based on its statistical distribution and the network status. Two aspects need to be considered in this step. On the one hand, in facial activity analysis, previous studies [15], [23] have shown that not all facial regions make equal contributions. As a matter of fact, only a sparse set of regions on faces contains important information and contributes to the final prediction [15], [23], [32]. On the other hand, different images have different statistical distributions, and therefore, the key region(s) in each image probably differ. In summary, the location search should be *selective* and *adaptive*.

To this end, we utilize the point adversarial attack in this step. As an adversarial attack method, the point adversarial attack can locate an image region that is sensitive to final predictions. The location searching process is adaptive since it is dependent on two factors: 1) the statistical distribution of the image itself and 2) prediction capability of the network used to search the location. Unlike previous unrestricted adversarial attack works such as [56], the point adversarial attack searches one solo position for each given image, which not only complies with the findings in [15], [23], and [64] but also benefits to the interpretability.

The general goal of adversary attack can be formulated as follows:

$$\begin{aligned} \max_{e(\mathbf{x})^*} \quad & F_{\mathrm{adv}}(\mathbf{x} + e(\mathbf{x}), w) \\ \text{s.t.} \quad & c(e(\mathbf{x})) \end{aligned} \tag{2}$$

where $F(*, w)$ denotes the target network, $\mathbf{x}$ denotes the original input without perturbations, $e(\mathbf{x})$ is an "additive adversarial perturbation" [59] with respect to $\mathbf{x}$, $F(\mathbf{x}, w)$ is the prediction for $\mathbf{x}$, and $F(\mathbf{x} + e(\mathbf{x}), w)$ is the prediction for the perturbed input, that is, $\mathbf{x}+e(\mathbf{x})$. An adversarial attack aims to find perturbations under a specified constraint, that is, $c(e(\mathbf{x}))$, to mislead the network $F(*, w)$. In our case, the constraint $c(e(\mathbf{x}))$ is $\|e(\mathbf{x})\|_0 \leq d$, in which $d$ is set as one to make the perturbation applied on one solo position.

We follow the point attack method proposed in [59] to locate the informative position in an adaptive manner. The key part of the point attack calculation is based on differential evolution, rather than gradient descent/ascent. Starting from a population of solution candidates, each of which denotes a perturbation and is encoded as a vector, a differential evolution method conducts population selection and inheritance to generate better solutions for the target. Concretely, each solution candidate, that is, $\hat{x}_*()$, encodes the spatial coordinates to apply the point attack. At each iteration (generation), a new candidate solution is produced by the following formulation:

$$\hat{x}_i(g+1) = \hat{x}_{r1}(g) + k\big(\hat{x}_{r2}(g) - \hat{x}_{r3}(g)\big) \tag{3}$$

where $g$ is the generation index, $\hat{x}_i$ is a candidate solution, $r1$, $r2$, and $r3$ are the candidate index in the same generation,[1] and $k$ is a predefined scale factor. A new candidate solution in generation $g+1$ is generated by three different solution candidates selected in generation $g$. The candidate solution will compete with their parents and it will be saved for further calculation only if it is better than its parents.

For each training sample, we use the point adversarial attack method discussed above to calculate an attack sample $\hat{x}_*()$, which contains the attack position denoted as $(p^x, p^y)$. The calculation process needs to consider the statistical information of the input and the prediction capability of network $F(*, w)$. For different samples, since they have different statistical information, the calculated attack positions for each image are different; for each sample, if we search the attack position using different networks, for example, $F(*, w)$ saved in different epochs, the located attack positions are also different. Therefore, the sensitive positions located in our method meet the original requirements: selective and adaptive.

### C. Generating New Learning Materials to Learn

In this step, we utilize the sensitive positions located in the previous step, that is, $(p^x, p^y)$ to generate learning materials for network refinement. The new learning materials are generated by erasing the local information centered at the located position $(p^x, p^y)$, since the perturbations added at those located positions are easy to mislead the network. This step is like providing new textbooks for students in each semester, based on the learning capability of students.

Concretely, for each given sample $\mathbf{x}_i$, assuming its size is $W \times H$, where $W$ denotes the width and $H$ denotes the height, and the sensitive position located in Section III-B is $(p^x, p^y)$, we mask out the information in a local region centered at $(p^x, p^y)$ with a size of $s \times s$. The generated samples are used as new learning materials to improve the capability of the network.

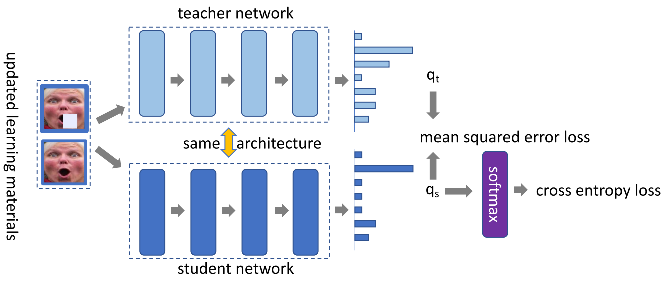


Fig. 2. Network refinement by updated learning samples and updated teachers. Unlike previous teacher–student works, in which the student network is with a smaller size, our teacher and student network share the same architecture. The deeper color, the better performance. Best viewed in color.

### D. Update New Teacher to Guide Student

To train a network with better generality, we not only utilize the new learning materials, denoted as $\mathbf{x}'$, but also a teacher network to provide guidance during the learning.

Specifically, as illustrated in Fig. 2, the network used to locate the attack information acts as a teacher network now, providing guidance to learn a new network, that is, a student network with the same architecture. We denote the student network as $F_s(*, w_s)$. The teacher network is fixed in the current round and provides guidance to train the student network. For each sample $\mathbf{x}_i$, the teacher network provides logits signal denoted as $q_i^t$. The logits $q_i^t$ generated by the teacher network, as well as the corresponding one-hot vector label $y_i$, are both utilized to supervise the student network training. Therefore, other than new learning samples, the student network receives additional guidance from the teacher network. To supervise the student network learning, we use a CE loss to minimize the discrepancy between the prediction and the ground-truth hard label, and a mean-squared error (MSE) loss to minimize the discrepancy between the logits from the teacher network and the student network. The formulation is as follows:

$$\begin{aligned} L = {} & \alpha * \mathrm{CE}\big(F_s\big(x_i', w_s\big), y_i\big) \\ & + \beta * \mathrm{MSE}\big(q_i^s\big(w_s, x_i'\big), q_i^t\big(w_t, x_i'\big)\big) \end{aligned} \tag{4}$$

where $F_s(*, w_s)$ denotes the student network with parameter $w_s$, $q_i^s(x_i', w_s)$ denotes the logits generated by the student network, while $q_i^t(x_i', w_t)$ denotes the logits generated by the teacher network, and $\alpha$ and $\beta$ are parameters to balance the contribution of the two loss terms.

There are two advantages in our teacher–student training strategy: 1) the updated learning samples and the teacher–student optimization assist the network to better model interclass variations in FER [65] and 2) unlike previous teacher–student learning works such as [66], our teacher network and student network share the same structure, relieving us from designing or selecting a proper teacher network and student network. We name our learning strategy as the self-mining mechanism.

*Iterative Mining Mechanism:* The aforementioned step can be conducted in an iterative way. In each iteration, the student network trained in the previous iteration changes its role, and becomes a teacher network in the current iteration. Since the network capability keeps improving as the iteration

[1] $r1 \neq r2 \neq r3$.

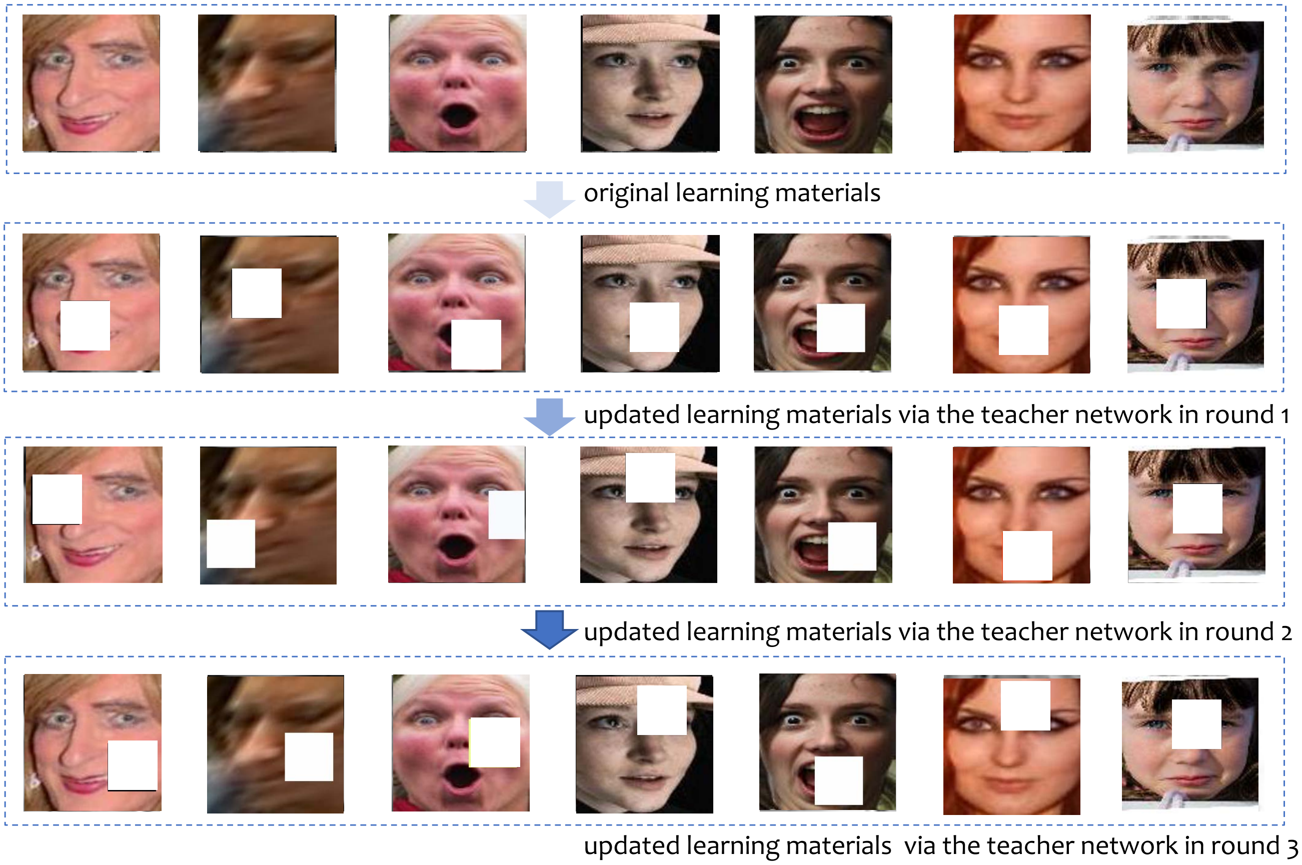

original learning materials
updated learning materials via the teacher network in round 1
updated learning materials via the teacher network in round 2
updated learning materials via the teacher network in round 3

Fig. 3. Visualization of updated learning images by PASM. Images in each row are new learning materials generated by an updated teacher network, which was a student in the previous iteration (row). Best viewed in color.

increases, the located sensitive position for each image is different and the generated new learning samples vary. The process is illustrated in Fig. 3. With the help of updated new learning materials and a stronger teacher network, the student network capability keeps improving correspondingly. The details of our algorithm can be found in Algorithm 1.

## IV. Experimental Results and Discussion

In this section, we utilize in-the-wild datasets, that is, real-world affective face database (RAF-DB) 2.0 [67], FER-2013 [68], Occlusion-RAF-DB [26], and Pose-RAF-DB [26], and one lab-controlled dataset, that is, Extended CohnKanade (CK+) [6] to demonstrate the efficacy of our method. All the details about these datasets and evaluation settings are described in the following sections.

### A. Experimental Setup

We test our method on three frequently used facial expression datasets, that is: 1) FER-2013 [68]; 2) RAF-DB 2.0 [67]; and 3) Extended CohnKanade (CK+) [6]. RAF-DB and FER-2013 are collected in the wild, covering variations in the real world, such as lighting conditions, large head pose, etc. To demonstrate the efficacy of our method when facing occlusion and pose issues in real scenarios, we further conduct experiments on Occlusion-RAF-DB [26] and Pose-RAF-DB [26], which are constructed subsets from RAF-DB with additional occlusion and pose annotations.

*RAF-DB* [67] is a dataset collected from the Internet. There are 29 672 facial images in this dataset. RAF-DB consists of highly diverse samples and covers different variations in the real world. The labels in this dataset are manually achieved by crowdsourced annotation [67]. This dataset has two kinds of annotations: 1) basic expressions and 2) compound expressions. To make comparisons with previous works, we only utilize the basic expression label set to test our method. For the basic expression label set, there are 12 271 images for training and 3068 images for testing.

*FER-2013* [68] is another widely used in-the-wild dataset. There are 28 709 samples in the training set and 3589 samples in the testing set. All those images are collected by the Google search engine and labeled with basic expressions. It was constructed for the ICML 2013 challenges in representation learning. It should be noted that the original image size in this dataset is only $48 \times 48$. The small spatial size and high variations make the analysis difficult.

*CK+* [6] is a dataset collected in a lab controlled environment. The original data collected is video based, which has 593 video sequences in total. In the first frames of each sequence, the subject activates a neutral expression and gradually shifts to a peak expression in the last frames. There are 327 sequences with basic expression labels in this dataset.

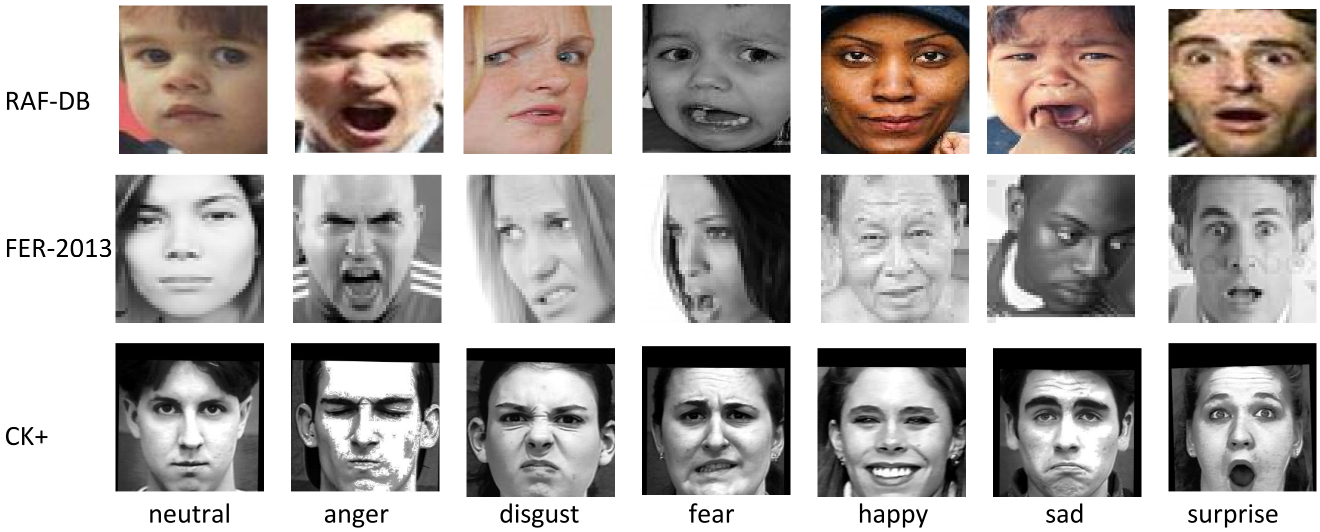


Fig. 4. Sample visualization for datasets utilized in this work, that is, RAF-DB, FER-2013, and CK+. From left to right: neutral, anger, disgust, fear, happy, sad, and surprise. From top to bottom: RAF-DB 2.0, FER-2013, and CK+. RAF-DB 2.0 and FER-2013 are in-the-wild databases, while CK+ is a dataset collected in controlled environments. Best viewed in color.

Unlike RAF-DB and FER-2013, there is no official training/validation/test split. Previous works usually utilize the first frame as a neutral sample and the last three frames with the target expression labels to construct the data. In the constructed data, there are 1308 images labeled with seven basic expressions in total.

*Occlusion-RAF-DB and Pose-RAF-DB* [26] are occlusion test subsets extracted from RAF-DB. In those subsets, there are various occlusion types, such as masks and glasses on faces, objects before faces, etc. Those constructed subsets are effective to test the efficacy of FER networks when facing real scenarios. Existing FER methods have a drop in recognition accuracy when testing in real cases full of occlusion and pose variations [26].

Examples from those three datasets are shown in Fig. 4. There are different variations in those datasets, including: 1) *scale:* the spatial size of each sample in FER-2013 is $48 \times 48$, which are much smaller than that in RAF-DB and CK+; 2) *lighting:* the lighting conditions when capturing those datasets are quite different; and 3) *pose:* there are heavy pose variations in RAF-DB and FER-2013, which are closer to the real scenarios.

To demonstrate the efficacy of our method, we utilize those aforementioned datasets for three evaluation settings, that is: 1) inner-database evaluation; 2) cross-database evaluation; and 3) occlusion-subset evaluation. The details of the three settings are illustrated as follows.

*Inner-Database Evaluation:* In inner-database evaluation, we train the network on the training set of a dataset and test the trained network on the testing set of the same dataset. We conduct an inner-database evaluation on RAF-DB and FER-2013.

*Cross-Database Evaluation:* In cross-database evaluation, we train the network on the training set of a dataset and test the trained network on a different dataset. To make a fair comparison with previous work [69], in this setting, we conduct the training on RAF-DB and test the trained network on FER-2013 and CK+.

*Occlusion and Pose Subset Evaluation:* In this setting, we train the network on RAF-DB by our method, and test it on the Occlusion-RAF-DB and Pose-RAF-DB, demonstrating the efficacy of our proposed method when dealing the challenging cases.

### B. Implementation Details

To preprocess the data, we detect faces and conduct the alignment based on MTCNN [70] from each given image. All the aligned detected faces are resized to $224 \times 224$. For making fair comparisons with previous works, we choose ResNet-34 and VGG-16 as our backbones, which are pretrained on the ImageNet dataset [71]. The output number for both networks is reset to seven, corresponding to the number of basic expressions. We optimize our networks by stochastic gradient descent with 0.9 momentum. We set the initial learning rate to 0.01, which will be multiplied by 0.1 every ten epochs. We implement all the experiments in PyTorch [72] and run our experiment on NVIDIA GTX 2080Ti GPU cards.

### C. Performance Evaluation

In this section, we test the discriminative ability of our method by conducting experiments for inner-dataset evaluations, and test the generality capability of our method by conducting experiments for cross-dataset evaluations. Our experiment is evaluated via mean classification accuracy.

*1) Inner-Dataset Evaluations:* In this section, we utilize two in-the-wild databases, that is: 1) RAF-DB and 2) FER-2013, to demonstrate the efficacy of our method. The result comparisons on the two datasets are reported in Tables I and II, respectively.

*Result Comparison on RAF-DB:* We compare our method to previous state-of-the-art methods on RAF-DB and report the

TABLE I
INNER-DATASET COMPARISON ON RAF-DB

| Method | Year | Backbone | Accuracy |
|---|---|---|---|
| FSN[74] | 2018 | AlexNet | 81.10% |
| MRE-CNN[75] | 2018 | VGG-16 | 82.63% |
| PAT-VGG-F-(gender,race)[73] | 2018 | VGG-16 | 83.83% |
| PAT-ResNet-(gender,race)[73] | 2018 | ResNet-34 | 84.19% |
| OADN[24] | 2020 | ResNet-50 | 87.16% |
| SCN[25] | 2020 | ResNet-18 | 87.03% |
| RAN[26] | 2020 | ResNet-18 | 86.90% |
| PASM | 2020 | VGG16 | **87.50**% |
| PASM | 2020 | ResNet-18 | **87.18**% |
| PASM (3 rounds) | 2020 | ResNet-34 | **88.68**% |

TABLE II
INNER-DATASET COMPARISON ON FER-2013

| Method | Year | Backbone | Accuracy |
|---|---|---|---|
| Guo et al.[76] | 2016 | InceptionNet | 71.33% |
| ECNN[77] | 2017 | Ensemble | 69.96% |
| Ron et al.[78] | 2017 | 3 Conv | 72.1% |
| PAT-VGG-F-(gender,race).[73] | 2018 | VGG-16 | 72.16% |
| PAT-ResNet-(gender,race).[73] | 2018 | ResNet-34 | 72.00% |
| PASM | 2020 | VGG-16 | **72.73**% |
| PASM (2 rounds) | 2020 | ResNet-34 | **73.59**% |

TABLE III
COMPARISON ON OCCLUSION-RAF-DB AND POSE-RAF-DB

| RAF-DB | Occlusion | Pose(30) | Pose(45) |
|---|---|---|---|
| RAN [26] | 82.72% | 86.74% | 85.20% |
| RanEra [21] | 78.78% | 85.00% | 83.69% |
| Our method | **83.27**% | **89.66**% | **88.17**% |

TABLE IV
CROSS-DATASET COMPARISON ON CK+

| Method | Source | Target | Accuracy |
|---|---|---|---|
| CNN-Li [69] | RAF-DB | CK+ | 78.00% |
| Our method | RAF-DB | CK+ | **79.65**% |

TABLE V
CROSS-DATASET COMPARISON ON FER-2013

| Method | Source | Target | Accuracy |
|---|---|---|---|
| CNN-Li [69] | RAF-DB | FER-2013 | 55.38% |
| Our method | RAF-DB | FER-2013 | **54.78**% |

comparison in Table I. In Table I, [73] introduces more manual labels into training and therefore, needs more label cost, and [24] and [26] design the region attention branch network, placing different weights on facial regions based on their occlusion conditions. Comparing to those previous works [24], [26], [73], our method achieves higher performance on both chosen architectures, that is, ResNet-34 and VGG-16. To fairly compare with the latest works [25], [26], we test our method on ResNet-18 and still outperform [25] and [26]. Unlike those previous works, our method does not modify architectures or introduce any external data, bringing few additional computational or memory cost in the inference stage.

*Result Comparison on FER-2013:* We compare our method to previous state-of-the-art methods on FER-2013 and report the comparison in Table II. The FER-2013 dataset was constructed for a FER challenge. The small size of each sample and high variations introduced by pose and light conditions make the prediction rate much lower than RAF-DB. To improve the prediction accuracy on this dataset, various methods have been proposed. For example, [73] introduces an attribute tree CNN to explicitly model the facial attributes, such as race, gender, and age. From Table II, we can find that our method achieves a higher or comparable accuracy on this challenging dataset.

*2) Evaluations on Occlusion-RAF-DB and Pose-RAF-DB:* Occlusion-RAF-DB and Pose-RAF-DB are two test subsets with occlusion and pose annotations. Those two datasets are constructed to test the network capability under occlusion and pose variations. Wang *et al.* [26] designed a region attention network (RAN), which adaptively assigns different weights to each facial region based on their contributions. Other than that, a region biased loss is proposed. As shown in Table III, our method, that is, PASM, outperforms [26] on all three subsets. Specifically, for the occlusion subset, the gain is 0.55%; for pose larger than 30° and 45°, the gains are 2.92% and 2.97%, respectively. Compared to random erasing [21], which does not have an *iterative* self-mining mechanism, PASM achieves higher performance on all three subsets. The better performance of our method on those three challenging datasets demonstrates that the better performance of our method comes from continuously updated new learning materials and advanced teacher networks.

*3) Cross-Dataset Evaluations:* To test the generality of our method, we conduct a cross-dataset evaluation. We train the network on RAF-DB and test it directly on CK+ and FER-2013. The comparisons with previous works are reported in Tables IV and V. FER-2013, as discussed in previous sections, is a challenging dataset collected in the real world; CK+ is collected in a lab-controlled situation and is the most representative in-the-lab dataset. We want to test the generality of our method on those two datasets. Again, by taking advantages of new learning materials and strong guidance provided from updated teacher networks, our method achieves a higher cross-dataset accuracy on CK+, a comparable accuracy on FER-2013.

### D. Ablation Study

We conduct extensive ablation studies in this section. At first, we analyze the generality of the self-mined knowledge; second, we conduct a comparison with the random erasing and adversarial erasing for demonstrating the superiority of our self-mined strategy; third, we analyze the relationship between the self-mining iteration number and the final accuracy; fourth, we illustrate the impact of the size of erasing mask and parameter in (4) to the final performance; last but not least, we conduct an evaluation of using random values for erasing.

*Generality of the Mined Knowledge:* We analyze the generality of self-mined knowledge extracted by our method and

TABLE VI
STUDY OF SELF-MINED INFORMATION GENERALITY

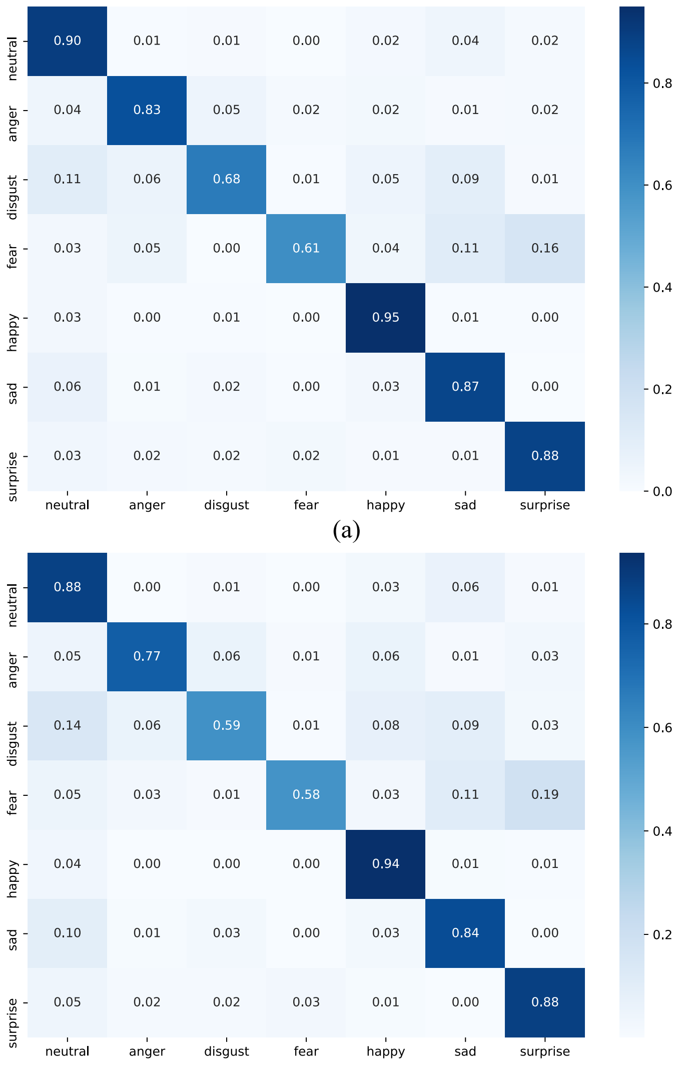
| Dataset | Mining Network | Training Network | Accuracy |
|---|---|---|---|
| PAT[73] | - | ResNet-34 | 83.83% |
| PAT[73] | - | VGG-16 | 84.19% |
| OADN[24] | - | ResNet-50 | 87.16% |
| PASM | ResNet-34 | ResNet-34 | **87.54**% |
| PASM | ResNet-34 | VGG-16 | **87.09**% |
| PAT[73] | - | ResNet-34 | 72.16% |
| PAT[73] | - | VGG-16 | 72.00% |
| PASM | ResNet-34 | ResNet-34 | **73.50**% |
| PASM | ResNet-34 | VGG-16 | **72.73**% |

report the results in Table VI. A "mining network" means the network used to locate a sensitive position to generate new learning samples, and the "training network" denotes the network for extracting knowledge from learning materials. As shown in Table VI, when the "training network" is different from the mining network, the prediction rate drops a bit on both datasets. On RAF-DB, the performance drops from 87.54% to 87.09%; and on FER-2013, the prediction rate drops from 73.50% to 72.73%. We believe this performance drop comes from the difference between the mining network and the training network, for example, network structures, parameters. However, even when we use two different networks for self-mining and training, our performance on RAF-DB and FER-2013 still outperforms or is comparable with previous works, that is, [24] and [73].

*Comparison With Random Erasing and Adversarial Erasing:* We make a comparison with the random erasing [21] and adversarial erasing [22]. Since random erasing [21] does not have an iterative self-mining mechanism, *for example, without updating learning materials or updating teacher networks working collaboratively*, we do not utilize teacher–student optimization and iterative refinement in this experiment (PASM w/o teacher). The comparison results are reported in Table VII. We can find that our method outperforms random erasing and adversarial erasing. Compared to random erasing conducting erasing in a random position, our method locates the position in a more adaptive manner by both considering the statistical information of each sample and the network capability. The lower performance of random erasing [21] compared to adversarial erasing [22] also demonstrates that an adaptive manner outperforms a random way. Compared to adversarial erasing, our method achieves better accuracy. We erase the subregion located in a more structured and sparse way, that is, a solo position, which aligns with the findings in previous works.

We show the confusion matrix for PASM and random erasing in Fig. 5. It can be observed that PASM outperforms random erasing especially on "anger" and "disgust."

*Analysis of Self-Mining Iterations:* In Fig. 6, we report the impact of self-mining iteration numbers in our proposed method. As illustrated in Section. III, PASM can be conducted in an iterative way. The student network in the current iteration changes its role to a teacher network in the following iteration, generating new studying materials and providing guidance to the new student network. We report the performances of a ResNet-34 on RAF-DB and FER-2013 in different iterations. Based on the experimental results, we can find that: 1) on the RAF-DB dataset, the accuracy rates increase with the changing iteration numbers, but the rate of increase decreases and 2) on FER-2013, the accuracy rates increase first; after two iteration refinements, the performance starts to drop (but still higher than the first round).



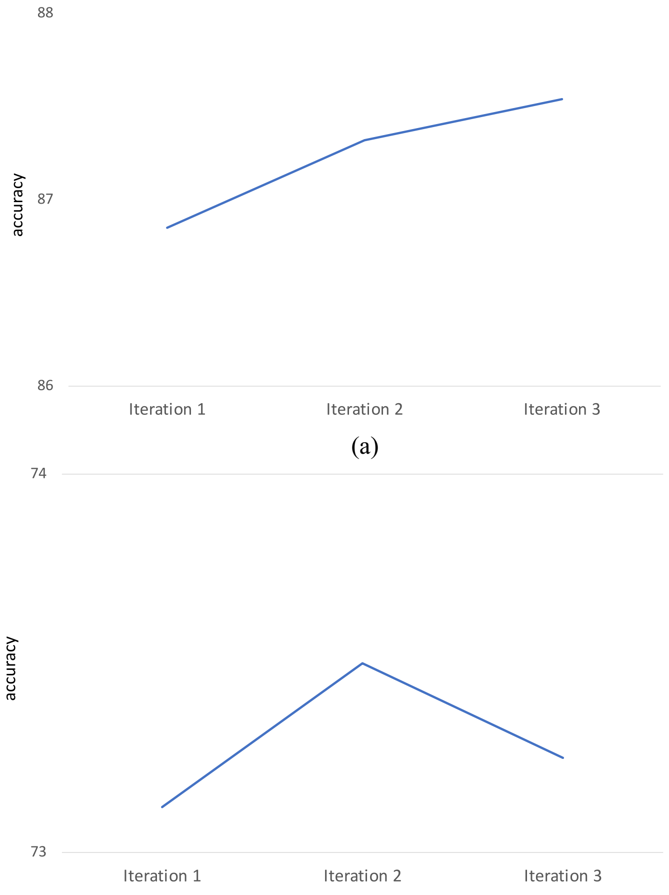
(a)

(b)

Fig. 5. Confusion matrix of (a) PASM and (b) random erasing on RAF-DB. The darker the color, the higher the accuracy.

*Analysis of Mask Sizes and Parameter Configuration in (4):* In Fig. 7, we report the prediction accuracy with different mask sizes, which is trained on RAF-DB by a ResNet-34 for one round. We choose five different mask sizes, that is, from 23 to 39 with a step size of 4. In the figure, we can find that the accuracy is not very sensitive to the mask sizes.

In Fig. 8, we show the prediction accuracy with parameter configuration in (4). In this experiment, we train on RAF-DB via a ResNet-34 for one round. We set $\alpha$ as 1, and change $\beta$ to 0.5, 0.1, and 0.01, respectively. In the figure, we can find that the accuracy is stable with different parameter configurations.



(a)

(b)

Fig. 6. Performance improvement of PASM with changing the iteration number. Based on the observations, we can find that: 1) on the RAF-DB dataset, the accuracy increases with the changing iteration numbers, but the rate of increase is decreasing and 2) on FER-2013, the accuracy increases first, and after two iterations of refinement, the performance starts to drop (but still higher than the first round). (a) ResNet-34 on RAF-DB. (b) ResNet-34 on FER-2013.

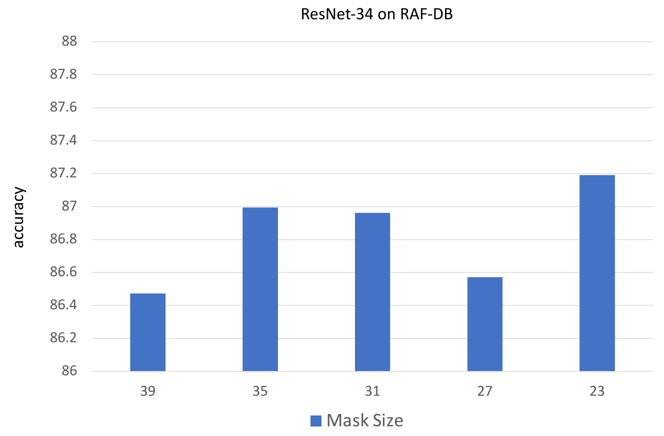


Fig. 7. Evaluation with different mask sizes. ResNet-34 trained on RAF-DB.

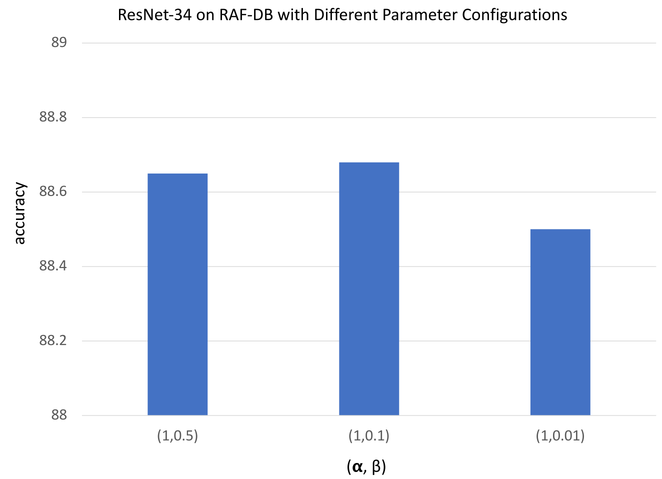


Fig. 8. Evaluation with different parameter configurations in (4). ResNet-34 trained on RAF-DB.

TABLE VII
COMPARISONS WITH RANDOM ERASING AND ADVERSARIAL ERASING. RESNET-34 IS CHOSEN AS THE BACKBONE AND TRAINED FOR ONLY ONE ROUND

| Dataset | Random Erasing | Adv Erasing | PASM w/o teacher |
|---|---|---|---|
| RAF-DB | 85.93% | 86.57% | **86.86%** |
| FER-2013 | 72.92% | 73.08% | **73.24%** |

*Evaluation of Random Erasing Values:* We conduct another experiment about using random values for erasing. Specifically, we generate random values between 0 and 255 and use the generated random values to replace the original pixels in the erasing region. The experiment is conducted on RAF-DB with a ResNet-34, trained for one round, without teacher guidance but with new learning samples. The accuracy is 86.93%, which is higher than random erasing (85.93%), comparable with PASM without teacher reported in Table VII (86.86%). This matches our expectation: 1) generating new learning materials is important to improve network capability and 2) when generating new learning materials to benefit PASM learning processes, how to locate the position adaptively is more important than how to process the located position.

*Open Problems in PASM:* There are two open problems that need to solve in the future: 1) the sensitive position search, as an adversarial attack in essential, might bring additional computation cost in training (only), while in inference stages, our method does not bring any additional parameters and computation cost. Currently, the point adversarial attach step consumes around 10 s for a 256×256 image. This computation cost issue in adversarial learning has already aroused research attention in recent works such as [79], and it should be mitigated by a more advanced adversarial attack solution in the future and 2) an automatic and adaptive way for choosing an appropriate iteration number in PASM is expected. We leave those to our future work.

## V. CONCLUSION

In this article, we proposed a self-mining framework called PASM to improve the performance of FER. By progressively generating new learning materials and providing guidance from an updated teacher network, the recognition capability of the student network with the same architecture can be improved in an iterative manner. The experimental results on benchmark facial expression datasets have demonstrated the efficacy of the proposed method. In the future, we plan to explore the possibility to apply our method on other facial activity analysis problems, such as the facial action unit analysis.

## References

[1] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.

[2] S. Wang, B. Pan, H. Chen, and Q. Ji, "Thermal augmented expression recognition," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2203–2214, Jul. 2018.

[3] X. Zhao, J. Zou, H. Li, E. Dellandréa, I. A. Kakadiaris, and L. Chen, "Automatic 2.5-D facial landmarking and emotion annotation for social interaction assistance," *IEEE Trans. Cybern.*, vol. 46, no. 9, pp. 2042–2055, Sep. 2016.

[4] J. Jang, H. Cho, J. Kim, J. Lee, and S. Yang, "Facial attribute recognition by recurrent learning with visual fixation," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 616–625, Feb. 2019.

[5] P. Rodriguez *et al.*, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Trans. Cybern.*, early access, Feb. 9, 2017, doi: 10.1109/TCYB.2017.2662199.

[6] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2010, pp. 94–101.

[7] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 454–461.

[8] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 699–714, May 2005.

[9] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 229–234.

[10] M. Eckhardt, I. Fasel, and J. Movellan, "Towards practical facial feature detection," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 3, pp. 379–400, 2009.

[11] P. Yang, Q. Liu, and D. N. Metaxas, "Boosting coded dynamic features for facial action units and facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–9.

[12] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2008, pp. 1–6.

[13] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based hog features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 884–888.

[14] S. Han, Z. Meng, P. Liu, and Y. Tong, "Facial grid transformation: A novel face registration approach for improving facial action unit recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 1415–1419.

[15] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1805–1812.

[16] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 558–565.

[17] W. Xie, L. Shen, and J. Duan, "Adaptive weighting of handcrafted feature losses for facial expression recognition," *IEEE Trans. Cybern.*, vol. 51, no. 5, pp. 2787–2800, May 2021.

[18] H. Zhang, W. Su, and Z. Wang, "Weakly supervised local-global attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 37976–37987, 2020.

[19] 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.

[20] Y. Liu, J. Peng, J. Zeng, and S. Shan. (2019). *Pose-Adaptive Hierarchical Attention Network for Facial Expression Recognition.* [Online]. Available: http://arxiv.org/abs/1905.10059

[21] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.

[22] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6488–6496.

[23] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2562–2569.

[24] H. Ding, P. Zhou, and R. Chellappa, "Occlusion-adaptive deep network for robust facial expression recognition," 2020. [Online]. Available: arxiv.abs/2005.06040.

[25] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6896–6905.

[26] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and ccclusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, Jan. 2020, doi: 10.1109/TIP.2019.2956143.

[27] 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.

[28] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[29] T. Zhang, "Facial expression recognition based on deep learning: A survey," in *Proc. Int. Conf. Intell. Interact. Syst. Appl.*, 2017, pp. 1–9.

[30] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 860–865.

[31] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 966–979, Aug. 2012.

[32] P. Liu, J. T. Zhou, I. W.-H. Tsang, Z. Meng, S. Han, and Y. Tong, "Feature disentangling machine—A novel approach of feature selection and disentangling in facial expression analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 151–166.

[33] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[34] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3359–3368.

[35] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2168–2177.

[36] B. Hu, Z. Zheng, P. Liu, W. Yang, and M. Ren, "Unsupervised eyeglasses removal in the wild," *IEEE Trans. Cybern.*, early access, Jun. 8, 2020, doi: 10.1109/TCYB.2020.2995496.

[37] Y. Zhang, I. W. Tsang, J. Li, P. Liu, X. Lu, and X. Yu, "Face hallucination with finishing touches," *IEEE Trans. Image Process.*, vol. 30, pp. 1728–1743, Jan. 2021, doi: 10.1109/TIP.2020.3046918.

[38] W. L. Zheng, W. Liu, Y. Lu, B. L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[39] Z. Meng, S. Han, P. Liu, and Y. Tong, "Improving speech related facial action unit recognition by audiovisual information fusion," *IEEE Trans. Cybern.*, vol. 49, no. 9, pp. 3293–3306, Sep. 2019.

[40] J. Cai *et al.*, "Feature-level and model-level audiovisual fusion for emotion recognition in the wild," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2019, pp. 1–5.

[41] L. Shan and D. Weihong, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015. [Online]. Available: https://www.robots.ox.ac.uk/∼vgg/publications/2015/Simonyan15/

[44] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://github.com/facebookresearch/mixup-cifar10

[45] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, p. 60, Jul. 2019.

[46] H. Fan, P. Liu, M. Xu, and Y. Yang, "Unsupervised visual representation learning via dual-level progressive similar instance selection," *IEEE Trans. Cybern.*, early access, Mar. 11, 2021, doi: 10.1109/TCYB.2021.3054978.

[47] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2507–2516.

[48] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 424–440.

[49] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6777–6786.

[50] P. Pan, P. Liu, Y. Yan, T. Yang, and Y. Yang, "Adversarial localized energy network for structured prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5347–5354.

[51] J. Zhou *et al.*, "Locality-aware crowd counting," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 3, 2021, doi: 10.1109/TPAMI.2021.3056518.

[52] Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Category-level adversarial adaptation for semantic segmentation using purified features," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 8, 2021, doi: 10.1109/TPAMI.2021.3064379.

[53] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Adversarial style mining for one-shot unsupervised domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–12.

[54] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, p. 6.

[55] S. Rao, D. Stutz, and B. Schiele, "Adversarial training against location-optimized adversarial patches," 2020. [Online]. Available: arxiv.abs/2005.02313

[56] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–6.

[57] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.

[58] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–6.

[59] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

[60] L. Jiang, X. Ma, S. Chen, J. Bailey, and Y.-G. Jiang, "Black-box adversarial attacks on video recognition models," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 12338–12345.

[61] N. Papernot, P. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ASIA CCS ACM Asia Conf. Comput. Commun. Security*, 2017, pp. 506–519.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[63] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," in *Proc. Int. J. Comput. Vis. (IJCV)*, 2015, pp. 1–6.

[64] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, "Multi-class heterogeneous domain adaptation," *J. Mach. Learn. Res.*, vol. 20, no. 57, pp. 1–31, 2019.

[65] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2138–2147.

[66] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learn. Represent. Learn. Workshop*, 2015, pp. 1–6.

[67] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[68] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. ICONIP*, 2013, pp. 117–124.

[69] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Trans. Affect. Comput.*, early access, Feb. 11, 2020, doi: 10.1109/TAFFC.2020.2973158.

[70] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[71] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[72] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[73] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Probabilistic attribute tree in convolutional neural networks for facial expression recognition," 2018. [Online]. Available: arxiv.abs/1812.07067

[74] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in CNNs for facial expression recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 317.

[75] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *Proc. Int. Conf. Artif. Neural Netw.*, 2018, pp. 84–94.

[76] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshop*, 2016, pp. 47–59.

[77] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang, and E. Xun, "Ensemble of deep neural networks with probability-based fusion for facial expression recognition," *Cogn. Comput.*, vol. 9, pp. 597–610, May 2017.

[78] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," 2017. [Online]. Available: arXiv:1705.01842.

[79] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–9.