Learn From All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition

Yuhang Zhang[0000−0003−4161−5020], Chengrui Wang[0000−0003−0618−0797], Xu Ling[0000−0002−3495−9434], and Weihong Deng[0000−0001−5952−6996]

Beijing University of Posts and Telecommunications, Beijing, China
{zyhzyh, crwang, lingxu, whdeng}@bupt.edu.cn

Abstract. Noisy label Facial Expression Recognition (FER) is more challenging than traditional noisy label classification tasks due to the inter-class similarity and the annotation ambiguity. Recent works mainly tackle this problem by filtering out large-loss samples. In this paper, we explore dealing with noisy labels from a new feature-learning perspective. We find that FER models remember noisy samples by focusing on a part of the features that can be considered related to the noisy labels instead of learning from the whole features that lead to the latent truth. Inspired by that, we propose a novel Erasing Attention Consistency (EAC) method to suppress the noisy samples during the training process automatically. Specifically, we first utilize the flip semantic consistency of facial images to design an imbalanced framework. We then randomly erase input images and use flip attention consistency to prevent the model from focusing on a part of the features. EAC significantly outperforms state-of-the-art noisy label FER methods and generalizes well to other tasks with a large number of classes like CIFAR100 and Tiny-ImageNet. The code is available at https://github.com/zyh-uaiaaaa/Erasing-Attention-Consistency.

Keywords: Noisy label learning, Facial expression recognition, Erasing attention consistency

1 Introduction

Facial Expression Recognition (FER) has wide applications in the real world, such as driver fragile detection, service robots, and human-computer interaction [35]. The most common paradigm for FER is the end-to-end supervised manner, whose performance largely relies on the massive high-quality annotated data. However, collecting large-scale datasets with fully precise annotations is usually expensive and time-consuming, sometimes even impossible. Furthermore, facial expression images have inherent inter-class similarity (all classes are human faces) and annotation ambiguity (some expression images are quite confusing), making noisy label FER more challenging than traditional noisy label classification tasks. On the other hand, it is well-known that deep neural networks have enough capacity to memorize large-scale data with even completely random labels, leading to poor performance in generalization [2, 19, 48]. Therefore, robust

FER with noisy labels has become an essential and challenging task in computer vision [4, 7, 9, 18, 35, 38, 47, 49, 50].

Mainstream noisy label FER methods can be mainly classified into two categories, sample selection and label ensembling. SCN [38] and RUL [50] can be viewed as sample selection methods, which learn more from clean samples and then relabel the noisy samples. SCN [38] uses a fully-connected layer to learn an importance weight for each sample and suppresses uncertain samples during the training phase. RUL [50] learns uncertainty weights through comparison between different samples. IPA2LT [35] and DMUE [35] are label ensembling methods, which provide several labels for a single sample to better mine the latent truth. IPA2LT [35] assigns each sample more than one labels with human annotations or model predictions while DMUE [35] uses a multi-branch model to better mine the latent distribution in the label space. All the aforementioned methods get good performances under noisy label FER while they still have defects. Specifically, sample selection methods are based on the small-loss assumption [2, 48], which might confuse hard samples and noisy samples as both of them have large loss values during the training process. Sample selection methods also need the noise rate, which is non-trivial in large-scale real-world datasets. Label ensembling methods provide different views of the same sample using several networks, similar to crowdsourcing in real FER applications. However, the extra information gain they bring might be noisy. Label ensembling methods might bring great computation overhead, making them less preferable in real applications. Thus, the noisy label FER problem demands better methods that do not need to know the noise rate or train several models to perform well.

In this paper, instead of following the traditional path to detect noisy samples according to their loss values and then suppress them, we view noisy label learning from a new feature-learning perspective and propose a novel framework to deal with all the aforementioned defects. We find that the FER model remembers noisy samples by focusing on a part of the features that can be considered related to the noisy labels, shown in Figure 1. The image in the first column is labeled as sad, while its latent truth is surprise. SCN [38] remembers this noisy sample by focusing on the frown feature which can be considered related to the noisy label of the sad expression. However, it neglects the open mouth feature, which is vital for the correct classification as an open mouth combined with a frown leads to the latent truth surprise instead of the noisy label sad. From the attention regions of the noisy samples, we conclude that the FER model only observes a part of the features that can be considered related to the noisy labels to remember noisy samples. It is intuitive as remembering noisy samples by focusing on a part of the features that can be considered related to the noisy labels does not contradict the other learned features from the clean samples. Inspired by this finding, we propose to deal with noisy label FER from a new feature-learning perspective. If the model can not focus on a part of the features and always learns from the whole features, then it cannot remember the noisy samples. Learning from the whole features from all training samples also means

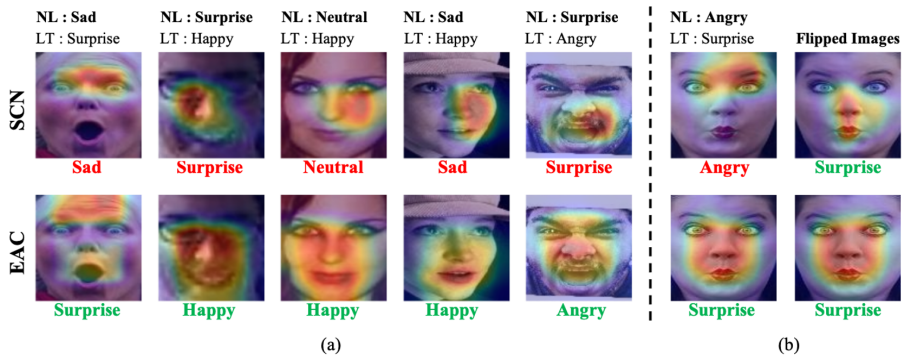


Fig. 1: (a) shows the attention regions of the noisy samples learned by SCN and EAC (Ours). NL represents the noisy label, LT represents the latent truth. The prediction results are shown under the images. SCN only focuses on a part of the features that can be considered related to the noisy labels to remember the noisy samples. (b) shows SCN predicts differently on the flipped image. Our EAC forces the model to focus on similar parts before and after the flip to prevent the model from remembering noisy labels.

the model does not need to filter out large-loss samples like traditional methods which might confuse useful hard samples with noisy samples.

In this paper, we use Attention Consistency to implement the consistency regularization. Attention Consistency [11] assumes that the learned attention maps should follow the same transformation as the input images to achieve better multi-label classification performance. The attention maps denote the features that the model based on to make the predictions.

We find that the flip semantic consistency of facial expression images can help to detect noisy labels. Flip semantic consistency means the original image and its flipped counterpart should be classified into the same category. However, if we train a FER model with a noisy sample, the model might remember the noisy sample while it still predicts the latent truth on its flipped counterpart, shown as the images in the first row of Figure 1. Inspired by that, we propose an imbalanced framework to prevent the model from remembering noisy samples. Specifically, we *only* compute classification loss on the original images and compute consistency loss between the attention maps extracted from the original images and their flipped counterparts. We utilize the consistency loss to prevent the model from remembering a part of the features of the original images. Such an imbalanced framework cannot help the model totally get rid of the noisy labels as the model can still gradually overfit the attention maps of the flipped images to keep the consistency loss small, which degrades the regularization effect. We further propose Erasing Attention Consistency (EAC) to increase the performance of the imbalanced framework. Before flipping, we first randomly erase the input images during the whole training phase. During the training

phase, the dynamic changing of the erased area ensures that the model can not simply remember the attention maps before and after the flip to get small consistency loss values. When the model starts to overfit the noisy original samples by focusing on a part of the features related to the noisy labels, the attention maps of the original images will deviate largely from the attention maps of their flipped counterparts, which will lead to large consistency loss values. We set the weight of the consistency loss larger enough to ensure the model first optimizes the consistency loss. Thus, to get small consistency loss values, the model will automatically quit overfitting the noisy samples.

The main contributions of our work are as follows:

1. Instead of using traditional methods which deal with noisy labels from high-level small-loss selection, we cope with noisy labels from middle-level feature learning, which does not require the noise rate to perform well.
2. We propose a novel method named Erasing Attention Consistency (EAC) which automatically prevents the model from memorizing noisy samples.
3. We experimentally show that EAC significantly advances state-of-the-art results on multiple FER benchmarks with different levels of label noise. EAC also generalizes well to image classification tasks with a large number of classes.

2 Related Work

Noisy Label Learning Learning with noisy labels has been well studied [1, 13–15, 17, 19, 20, 24, 25, 29, 31–33, 37, 41–43, 45, 46, 51]. Current works can be mainly categorized into two groups: modifying the primary loss function or selecting clean samples for training.

The first type of method mainly focuses on estimating the noise transition matrix or proposing robust loss functions. Patrini *et al.* [32] estimate the transition matrix to model the relationship between noisy labels and the latent truth to prevent the model from overfitting noisy labels. Han *et al.* [13] propose a human-assisted approach that conveys human cognition of invalid class transitions to make estimating transition matrix easier. Both Thulasidasan *et al.* [37] and Zhang *et al.* [51] propose generalized cross-entropy loss functions to combat noisy labels. Xu *et al.* [43] design a new loss function based on mutual information which is information-monotone and robust to various kinds of label noise. Although these methods have theory guarantees, they are not suitable for challenging real-world settings or handling a large number of classes. Thus, recent works usually focus on the second type of method.

The second strand of approach is based on the memorization effect that DNNs fit the underlying clean distribution before overfitting the noisy labels [2]. They focus on reweighting or sample selection to suppress noisy samples. Jiang *et al.* [19] train a mentor net using clean samples to guide the student net by weighing the samples. Ren *et al.* [33] reweight samples according to their gradient directions. Arazo *et al.* [1] model per-sample loss by a mixture model to calculate

a weight for each sample. Han *et al.* [14] train two models to select small loss samples for each other hoping to filter different types of error introduced by noisy labels. Malach *et al.* [29] improve co-teaching by updating only on instances with different predictions to keep the two models diverged. Wei *et al.* [41] train two models together and use their agreement degree to select small-loss samples. These methods select small-loss samples to eliminate the bad influence from the noisy samples. However, the useful hard samples are likely to have large loss values and might be filtered out as noisy samples. These methods also need to know the noise rate to get better performance. Different from them, our method automatically prevents the model from memorizing the noisy samples, which do not require the noise rate or selecting clean samples.

Facial Expression Recognition Facial Expression Recognition (FER) aims at helping computers to understand human behavior or even interact with a human by recognizing human expression. In recent years, as the recognition accuracy is very high in the laboratory collected FER datasets, more attempts try to address the in-the-wild FER problem, which contains lots of label noise. Zeng *et al.* [47] first consider annotation inconsistency and assign each sample with more than one label to better mine the latent truth. Wang *et al.* [38] propose to learn an importance weight for each sample and suppress the uncertain images by relabeling. She *et al.* [35] train multi-branch models by leaving out one class for each branch in order to find the latent truth under label noise. Zhang *et al.* [50] propose to learn the uncertainty of different facial images by comparison and then suppress the uncertain images. They can be mainly categorized into two classes, sample selection [38, 50] or label ensembling [35, 47]. Sample selection methods select good samples and suppress noisy samples while label ensembling methods use crowdsourcing to improve performance. However, they either require the noise rate to better filter out noisy samples or bring extra computation overhead and cannot generalize well to classification tasks with a large number of classes. Our method automatically prevents the model from overfitting the noisy samples without the noise rate and generalizes well to classification tasks with a large number of classes.

3 Proposed Method

In this section, we illustrate the implementation details of our proposed Erasing Attention Consistency (EAC) method.

3.1 Preliminary

Class Activation Mapping Class Activation Mapping (CAM) [53] is an attention method, which allows us to visualize the predicted class scores on the given images, highlighting the discriminative parts detected by the CNN.

In the CNN trained for classification, an attention map is the weighted sum of the feature maps from the last convolutional layer with the weights from a fully connected (FC) layer. By viewing the attention maps, we can know what

the model is based on to make the predictions. We denote the feature map extracted from the last convolutional layer as $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, C, H, W respectively represent the number of channels, height, width of the feature map. We denote the weights of the FC layer as $\mathbf{W} \in \mathbb{R}^{L\times C}$, L represents the number of classes. The attention map computes as

$$\mathbf{M}_j(h,w) = \sum_{c=1}^{C} \mathbf{W}(j,c)\mathbf{F}_c(h,w), \tag{1}$$

$\mathbf{M}_j(h,w)$ is the attention value of location (h,w) for class index j, which is the weighted sum of feature maps over different channels. In our method, we use CAM to compute the attention maps from the input images to show the features that the model attends to.

Attention Consistency Attention Consistency [11] is first proposed for achieving better visual perceptual plausibility and better multi-label image classification by considering visual attention consistency under spatial transforms. It assumes that the learned attention maps of the model should follow the same transformation as the input images.

3.2 Overview of Erasing Attention Consistency

In this paper, we design an imbalanced framework to help the model get rid of the negative effect of the noisy labels. We notice that the facial images before and after the flip have the same semantic meaning of the facial expression. We only compute classification loss with the original images and compute consistency loss between the attention maps of the original images and their flipped counterparts to prevent the model from remembering the original images with noisy labels. Simply using this imbalanced framework can not help the model totally get rid of the negative effect from noisy labels as the model can gradually remember the flipped images to always get small consistency loss, which degrades the regularization effect. We further propose Erasing Attention Consistency to enhance the performance of our proposed imbalanced framework. Before flipping the original images to generate their counterparts, we first randomly erase the images according to [52], which will generate different pairs of original images and their flipped counterparts during the training process. Thus, the model cannot remember the flipped images to get small consistency loss. If the model starts to remember the original images with noisy labels, the attention maps extracted from them will focus on a part of the features, which deviate largely from the flipped attention maps extracted from their flipped counterparts leading to the increase of the consistency loss. Thus, the consistency loss can prevent the model from remembering noisy samples.

3.3 Framework of Erasing Attention Consistency

The overall framework of our proposed EAC is shown in Figure 2. Given a batch of facial expression images, we first erase the input images according to

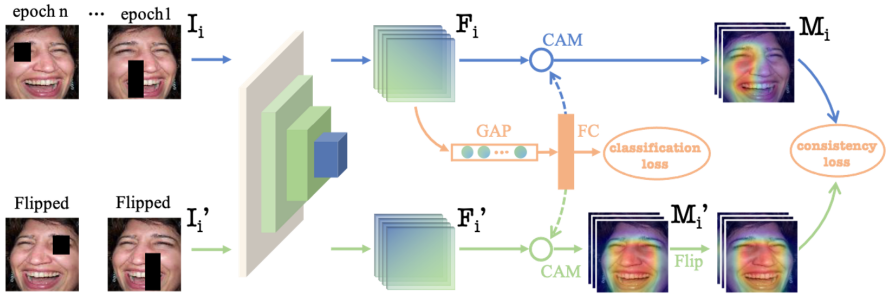


Fig. 2: The framework of the Erasing Attention Consistency (EAC). EAC randomly erases input images and then gets their flipped counterparts. EAC only computes the classification loss with the original images. The classification loss with the noisy labels might cause the model to overfit the noisy samples shown as M_i. EAC uses the consistency loss between the original images and their flipped counterparts to prevent the model from remembering noisy labels. The dotted lines mean no gradient propagation.

[52] and get $\mathbf{I}$. We then flip these images to get their flipped counterparts $\mathbf{I}^{'}$. $\mathbf{I}$ and $\mathbf{I}^{'}$ are the input images. The feature maps are extracted from the last convolutional layer, denoted as $\mathbf{F} \in \mathbb{R}^{N\times C\times H\times W}$ and $\mathbf{F}^{'} \in \mathbb{R}^{N\times C\times H\times W}$. N, C, H, W respectively represent the number of images, the number of channels, height, width of the feature maps. We *only* input $\mathbf{F}$ through the global average pooling (GAP) layer to get features $\mathbf{f} \in \mathbb{R}^{N\times C\times 1\times 1}$. We resize features $\mathbf{f}$ to $N\times C$ and put them through fully connected (FC) layer to compute classification loss according to

$$l_{cls} = -\frac{1}{N}\sum_{i=1}^{N}(\log\frac{e^{\mathbf{W}_{\mathbf{y}_i}\mathbf{f}_i}}{\sum_j^L e^{\mathbf{W}_j\mathbf{f}_i}}), \tag{2}$$

$\mathbf{W}_{\mathbf{y}_i}$ is the $\mathbf{y}_i$-th weight from the FC layer with $\mathbf{y}_i$ as the given label of the i-th image. We compute attention maps $\mathbf{M}$ and $\mathbf{M}^{'}$ for $\mathbf{I}$ and $\mathbf{I}^{'}$ according to Eq. (1). Note that the weights used to compute attention maps come from the FC layer, while the FC layer only computes classification loss with the original feature maps $\mathbf{F}$. We use consistency loss to minimize the distance between the feature maps $\mathbf{M}$ and $Flip(\mathbf{M}^{'})$ as

$$l_c = \frac{1}{NLHW}\sum_{i=1}^{N}\sum_{j=1}^{L}||\mathbf{M}_{ij} - Flip(\mathbf{M}^{'})_{ij}||_2. \tag{3}$$

The total loss is computed as follows,

$$l_{total} = l_{cls} + \lambda l_c. \tag{4}$$

λ is the weight of the erasing consistency loss. The ablation study of λ is in Section 4.8.

4 Experiments

In this section, we first describe 3 popular in-the-wild FER benchmarks and our implementation details. We then verify the proposed EAC on the FER datasets with different levels of label noise and study why EAC works. Visualization results of the learned features, attention maps and classification loss values are displayed to provide an intuitive understanding of EAC. We carry out an ablation study and also show the generalization ability of EAC by conducting experiments on CIFAR100 [22] and Tiny-ImageNet [34]. Finally, we compare EAC with other state-of-the-art FER methods.

4.1 Datasets

RAF-DB [26] is annotated with basic or compound expressions by 40 trained human coders. In our experiments, images with seven basic expressions (i.e. neutral, happy, surprise, sad, angry, disgust, fear) are used including 12,271 images for training and 3,068 images for testing.

FERPlus [3] is extended from FER2013 [10] with finer label annotations. It is collected by the Google search engine consisting of 28,709 training images and 3,589 test images. We use the most voting category as the annotation for a fair comparison [3, 38, 39].

AffectNet [30] is by far the largest FER dataset, which is collected from the Internet by querying expression-related keywords in three search engines containing more than one million images. There are 286,564 training images and 4,000 test images manually labeled to eight classes.

4.2 Implementation Details

By default, we use ResNet-18 [16] pre-trained on MS-Celeb-1M [12] as the backbone network with the same routine as [35, 38, 39, 50] for fair comparisons. The facial images are aligned and cropped with three landmarks [40], resized to 224×224 pixels. We only use the horizontal flip and the random erasing without any other data augmentation tricks to evaluate the effectiveness of our proposed method. During training, the batch size is 256. The initial learning rate is 0.0002. We use Adam [21] optimizer with weight decay of 0.0001 and ExponentialLR [27] learning rate scheduler with the gamma of 0.9 to decrease the learning rate after each epoch. The training ends at epoch 60.

4.3 Evaluation of EAC on Noisy FER Datasets

We quantitatively evaluate the improvement of our proposed EAC against other state-of-the-art noisy label FER methods. We explore the robustness of EAC with three levels of label noise including the ratio of 10%, 20%, 30% on RAF-DB, FERPlus, and AffectNet datasets. We follow [35, 38, 50] to generate noisy labels. As the generation of label noise is random, we re-implement other state-of-the-art methods on our generated noisy datasets to make fair comparisons with

Table 1: Evaluation of EAC on noisy FER datasets. We re-implement other state-of-the-art methods and test all the methods with the same noisy datasets to make fair comparisons. Results are computed as the mean of the accuracy from the last 5 epochs

Method	Noise(%)	RAF-DB(%)	FERPlus(%)	AffectNet(%)
Baseline	10	81.01	83.29	57.24
SCN (CVPR20)	10	82.15	84.99	58.60
RUL (NeurIPS21)	10	86.17	86.93	60.54
EAC (Ours)	10	88.02	87.03	61.11
Baseline	20	77.98	82.34	55.89
SCN (CVPR20)	20	79.79	83.35	57.51
RUL (NeurIPS21)	20	84.32	85.05	59.01
EAC (Ours)	20	86.05	86.07	60.29
Baseline	30	75.50	79.77	52.16
SCN (CVPR20)	30	77.45	82.20	54.60
RUL (NeurIPS21)	30	82.06	83.90	56.93
EAC (Ours)	30	84.42	85.44	58.91

them. We also consider the influence of the different backbones and backbones with or without pretraining.

Shown in Table 1, our method outperforms other state-of-the-art FER noisy label learning methods by a large margin. For example, EAC outperforms SCN under 30% label noise by 6.97%, 3.24%, 4.31% on RAF-DB, FERPlus, AffectNet respectively.

Note that, unlike SCN [38] and RUL [50], EAC does not need to modify the labels of the training samples. Relabeling has the risk of changing right labels to wrong labels, which is less flexible than our method as EAC can automatically learn useful information from all training samples. EAC does not need to know the noise rate or tell apart hard samples and noisy samples, which fundamentally solves the defects of sample selection methods as sample selection methods require the noise rate to filter out large-loss samples, which might contain useful hard samples and useless noisy samples.

We also study EAC with different backbones. With different backbones, λ is set to 5 under 0 and 10% noise, 10 under 20% and 30% noise. As shown in Table 2, adding EAC to MobileNet or ResNet-50 can both improve their performance. Baselines are also trained with erase and flip for a fair comparison. EAC achieves better results in all settings using ResNet-50 as backbone compared with ResNet-18 in Table 1. The experiments of EAC using an unpretrained model as backbone are shown in the supplementary material.

4.4 Why EAC works

We evaluate the three modules of the proposed EAC to find why EAC works well under label noise. The experiment results are shown in Table 3. Several

Table 2: The influence of different backbones on EAC. We carry out experiments on RAF-DB. Results are computed as the mean of the accuracy from the last 5 epochs

Method	0 noise	10% noise	20% noise	30% noise
MobileNet	83.31%	77.80%	70.60%	62.48%
MobileNet + EAC	86.47%	82.63%	81.65%	79.82%
ResNet-50	88.75%	83.44%	79.11%	71.67%
ResNet-50 + EAC	90.35%	88.62%	87.35%	85.27%

observations are concluded as follows. Without the flip attention consistency module, the model can not use the same semantic meaning from the flipped counterparts to regularize the classification loss, which is shown in the second row. Without the erasing, the model will gradually remember the attention maps from the flipped images to get small consistency loss values, which degrades the regularization effect. Without the imbalanced framework, the noisy labels will affect the images before and after the flip together. The model can remember the noisy samples before and after the flip together, making the consistency loss useless. However, when we combine the three modules, the performance skyrockets.

We believe it is the dynamic erasing that prevents the model from remembering the attention maps. Thus, the model needs to learn flip consistent features to minimize the consistency loss. As we only compute the classification loss with the original images (the imbalanced framework), if the model tries to remember the noisy samples, the features learned from these samples will deviate largely from their flipped counterparts, making the consistency loss large. As we set the weight of the consistency loss large enough, the model will first minimize the consistency loss. Thus, it will quit remembering the noisy samples.

4.5 Whether flip and erase is sufficiently valid for EAC

We use flip because we need *spatial transforms* to enable attention consistency following [11]. Other spatial transforms like Rotate or Scale are not very effective

Table 3: Evaluation of the three modules of EAC on RAF-DB with 30% label noise

flip attention consistency	imbalanced framework	erasing	RAF-DB
×	×	×	75.50
×	✓	✓	78.10
✓	×	✓	78.29
✓	✓	×	76.26
✓	✓	✓	84.42

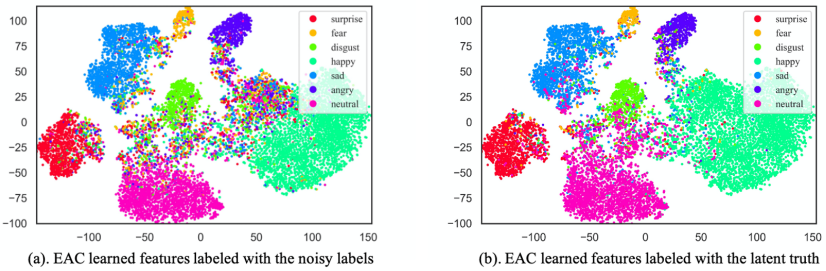
Table 4: Comparison with other augmentation methods. The experiments are carried out on noisy RAF-DB.

Noise	Rotate	Scale	Flip	Blur	AutoAug.	Erasing
10%	80.93%	85.98%	88.02%	86.80%	87.84%	88.02%
20%	79.63%	85.30%	86.05%	83.77%	85.82%	86.05%
30%	78.23%	82.01%	84.42%	76.92%	82.40%	84.42%

for FER as FER test sets are mainly frontal faces with a similar scale. We utilize erasing as FER models fit noisy labels through remembering parts of the features. Erasing guides the model to focus on the whole feature as the remembered feature parts might be absent during the training. Other augments can not directly solve the part-view problem and are not very effective. We test them on noisy RAF-DB. Rotate and Scale is compared to Flip. Blur [36] and AutoAugment [5] (AutoAug.) is compared to Erasing. AutoAugment searches and combines many kinds of augments together while it is still inferior to erasing.

4.6 Feature Visualization

To understand EAC intuitively, we plot the learned features of EAC trained with 30% noisy labels on RAF-DB by t-SNE [28]. Figure 3 (a) is the learned features displayed with the noisy training labels. It is shown that EAC does not remember noisy labels as features with different labels are clustered together. It is shown that the features with noisy labels are close to the classification boundary which means these samples are with large classification loss values. Thus, EAC separates clean and noisy samples effectively. We also plot the same



(a). EAC learned features labeled with the noisy labels

(b). EAC learned features labeled with the latent truth

Fig. 3: The learned features by EAC training with noisy labels. (a) is the learned features displayed with the noisy training labels, EAC does not overfit noisy labels as different classes mixed with each other. *Notice that noisy samples are pushed to the classification boundary by EAC.* (b) is the same learned features with (a), but displayed with the latent truth. Though we train EAC with noisy labels, it can still learn useful features related to the latent truth.

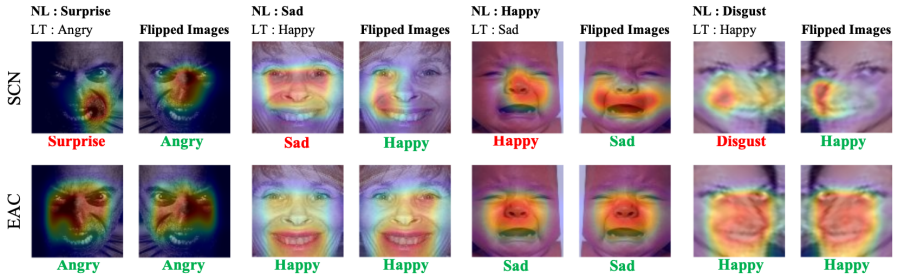


Fig. 4: The attention maps of SCN and EAC on the original images and their flipped counterparts.

learned features in Figure 3 (b), but displayed with the latent truth. Compared with Figure 3 (a), we can draw the conclusion that EAC can automatically prevent the model from remembering noisy labels and learn useful features from both clean and noisy samples.

We plot the attention maps on images before and after the flip in Figure 4 to show the effectiveness of EAC. We train SCN with the original images and test on their flipped counterparts. It is shown that SCN remembers the original images to the noisy labels, while it still gets correct predictions on their flipped counterparts after training. Inspired by that, EAC uses the attention maps of the flipped ones to regularize the classification loss and get correct prdictions on both the original images and their flipped counterparts. We display more results in the supplementary material.

4.7 Visualization of the classification loss values

We plot the distribution of classification loss values after training for 60 epochs in Figure 5 under the same setting as Section 4.6. We normalize the histogram of loss values and plot it as the probability density. The baseline method overfits

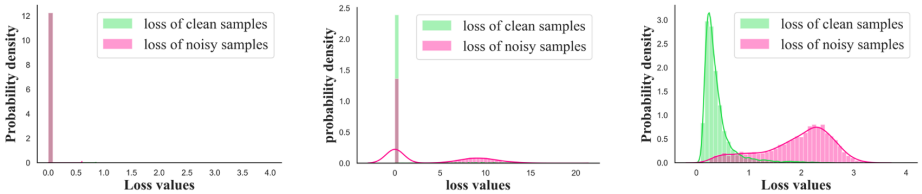


Fig. 5: The classification loss values of different methods after training for 60 epochs with noisy samples. The baseline remembers nearly all noisy samples. SCN avoids overfitting a part of the noisy samples, while EAC can still separate clean and noisy samples apart after training for 60 epochs.

nearly all the noisy samples after training for 60 epochs as the loss values of all samples are around 0. SCN learns importance weights and uses relabeling to deal with noisy samples. However, lots of the noisy samples are not correctly relabeled during the training process as there are still lots of noisy samples with loss values close to 0. Our EAC prevents the model from remembering the noisy samples during the whole training process. After training for 60 epochs, the loss values of clean and noisy samples can still be separated clearly.

4.8 Ablation Study

We evaluate the consistency loss weight λ from 0.1 to 10.0 with different levels of label noise. The results are shown in the supplementary material. We can choose λ from a wide range to acquire state-of-the-art performance. The best value of λ is 3 under 10% and 20% noise and 5 under 30% noise on RAF-DB using ResNet-18 as backbone. For simplicity, we set λ as 5 in the noisy label experiments using ResNet-18 as backbone.

4.9 The generalization ability of EAC

Noisy label FER methods might not be suitable for noisy label classification tasks with a large number of classes as the class number of the facial expression is very small. For example, DMUE [35] needs to train a multi-branch model whose branch number equals the class number plus 1 to mine the latent truth, which is unaffordable when the class number is very large. However, EAC can generalize well to tasks with a large number of classes.

To show the generalization ability of EAC. We carry out experiments on CIFAR100 [22] and Tiny-ImageNet [34]. Due to the space limitation, the implementation details are illustrated in the supplementary material. As shown in Tabel 5, our EAC consistently improves the baseline by a large margin in both top-1 and top-5 accuracy. EAC outperforms the baseline by 6.37% , 9.40%, 10.89% on CIFAR100 and 12.11%, 17.67%, 22.19% on Tiny-ImageNet in top-1 accuracy with noise ratio 10%, 20%, 30%. Although SCN [38] also outperforms the baseline, it is clear that our EAC achieves much better results.

Table 5: CIFAR100 and Tiny-ImageNet label noise training

Methods	CIFAR100 Noise Rate			Tiny-ImageNet Noise Rate		
	Top-1/Top-5 (%)			Top-1/Top-5 (%)		
	10%	20%	30%	10%	20%	30%
Baseline	64.56/85.37	57.33/78.93	49.70/72.55	58.11/80.24	49.56/72.43	41.32/64.58
SCN [38]	65.18/86.60	60.38/82.11	56.19/78.30	62.22/85.89	55.23/80.21	47.39/72.56
EAC	70.93/90.15	66.73/87.01	60.59/82.84	70.22/90.23	67.23/89.01	63.51/87.18

Table 6: Comparison with other state-of-the-art results on different FER datasets. † denotes training with both AffectNet and RAF-DB datasets. ∗ denotes test with 7 classes on AffectNet.

RAF-DB		FERPlus		AffectNet	
Methods	Acc. (%)	Methods	Acc. (%)	Methods	Acc. (%)
IPA2LT† [47]	86.77	IPA2LT† [47]	-	IPA2LT† [47]	57.31
RAN [39]	86.90	RAN [39]	88.55	RAN [39]	59.50
SCN [38]	87.03	SCN [38]	88.01	SCN [38]	60.23
DACL [8]	87.78	DACL [8]	-	DACL∗ [8]	65.20
KTN [23]	88.07	KTN [23]	90.49	KTN∗ [23]	63.97
DMUE [35]	88.76	DMUE [35]	88.64	DMUE [35]	62.84
RUL [50]	88.98	RUL [50]	88.75	RUL [50]	61.43
EAC (Ours)	89.99	EAC (Ours)	89.64	EAC∗(Ours)	65.32

4.10 Comparison with other state-of-the-art FER methods

EAC can also help the FER model achieve state-of-the-art performance on clean datasets as EAC encourages the model to learn flip consistent features from the input images which conforms to the human visual perceptual. The results are shown in Table 6. Besides the works mentioned in Section 2, RAN [39] utilizes attention weights to aggregate a varied number of face regions to recognize facial expression robustly. DACL [8] adaptively selects a subset of significant feature elements for enhanced discrimination. [23] utilizes a knowledgeable teacher network (KTN) and a self-taught student network (STSN) to transfer knowledge. Our EAC achieves the best performance than other state-of-the-art methods on RAF-DB and AffectNet(7 classes) while slightly lower than KTN [23] under FERPlus. We do not compare with [44] as it utilizes Vision Transformer [6] as backbone while we use ResNet-18 [16].

5 Conclusion

In this paper, we explore to deal with noisy label FER from a new feature-learning perspective and propose a novel and effective method named Erasing Attention Consistency (EAC). We design an imbalanced framework to utilize the erasing and flip consistency loss to prevent the model from remembering noisy labels. EAC does not require the noise rate or label ensembling. Extensive experiments verify that EAC outperforms other state-of-the-art noisy label FER methods on clean and noisy datasets. Furthermore, EAC generalizes well to noisy label classification tasks with a large number of classes.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China under Grant 62192784 and Grant 61871052.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: ICML (2019)
2. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: ICML (2017)
3. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ICMI (2016)
4. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y.: Label distribution learning on auxiliary label space graphs for facial expression recognition. In: CVPR (2020)
5. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Fan, X., Deng, Z., Wang, K., Peng, X., Qiao, Y.: Learning discriminative representation for facial expression recognition from uncertainties. In: ICIP (2020)
8. Farzaneh, A.H., Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: WACV (2021)
9. Gera, D., Balasubramanian, S.: Noisy annotations robust consensual collaborative affect expression recognition. In: ICCV (2021)
10. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: ICONIP (2013)
11. Guo, H., Zheng, K., Fan, X., Yu, H., Wang, S.: Visual attention consistency under image transforms for multi-label image classification. In: CVPR (2019)
12. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV (2016)
13. Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., Sugiyama, M.: Masking: A new perspective of noisy supervision. In: NIPS (2018)
14. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: NIPS (2018)
15. Han, J., Luo, P., Wang, X.: Deep self-learning from noisy labels. In: ICCV (2019)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
17. Huang, J., Qu, L., Jia, R., Zhao, B.: O2u-net: A simple noisy label detection approach for deep neural networks. In: ICCV (2019)
18. Jiang, J., Deng, W.: Boosting facial expression recognition by a semi-supervised progressive teacher. IEEE Transactions on Affective Computing (2021)
19. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: ICML (2018)
20. Kim, Y., Yim, J., Yun, J., Kim, J.: Nlnl: Negative learning for noisy labels. In: ICCV (2019)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

22. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech Report (2009)
23. Li, H., Wang, N., Ding, X., Yang, X., Gao, X.: Adaptively learning facial expression representation via cf labels and distillation. TIP (2021)
24. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
25. Li, J., Xiong, C., Hoi, S.C.: Learning from noisy data with robust representation learning. In: ICCV (2021)
26. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR (2017)
27. Li, Z., Arora, S.: An exponential learning rate schedule for deep learning. arXiv preprint arXiv:1910.07454 (2019)
28. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)
29. Malach, E., Shalev-Shwartz, S.: Decoupling" when to update" from" how to update". In: NIPS (2017)
30. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing (2017)
31. Nguyen, D.T., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Beggel, L., Brox, T.: Self: Learning to filter noisy labels with self-ensembling. arXiv preprint arXiv:1910.01842 (2019)
32. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: CVPR (2017)
33. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: ICML (2018)
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
35. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: CVPR (2021)
36. Shi, Y., Yu, X., Sohn, K., Chandraker, M., Jain, A.K.: Towards universal representation learning for deep face recognition. In: CVPR (2020)
37. Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., Mohd-Yusof, J.: Combating label noise in deep learning using abstention. arXiv preprint arXiv:1905.10964 (2019)
38. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: CVPR (2020)
39. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing (2020)
40. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: ICCV (2019)
41. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: CVPR (2020)
42. Xie, M.K., Huang, S.J.: Partial multi-label learning with noisy label identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
43. Xu, Y., Cao, P., Kong, Y., Wang, Y.: L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In: NIPS (2019)

44. Xue, F., Wang, Q., Guo, G.: Transfer: Learning relation-aware facial expression representations with transformers. In: ICCV (2021)
45. Ye, M., Yuen, P.C.: Purifynet: A robust person re-identification model with noisy labels. TIFS (2020)
46. Yi, K., Wu, J.: Probabilistic end-to-end noise correction for learning with noisy labels. In: CVPR (2019)
47. Zeng, J., Shan, S., Chen, X.: Facial expression recognition with inconsistently annotated datasets. In: ECCV (2018)
48. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. In: ICLR (2017)
49. Zhang, F., Xu, M., Xu, C.: Weakly-supervised facial expression recognition in the wild with noisy data. IEEE Transactions on Multimedia (2021)
50. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. In: NIPS (2021)
51. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: NIPS (2018)
52. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)
53. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)