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

Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng

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

## 1 Evaluation on noisy datasets without pretraining

Without pretraining, EAC also increases the performance of SCN by 5.77%, 5.36%, 6.71% under 30% label noise in the 3 Facial Expression Recognition(FER) datasets, which illustrates that EAC can still get rid of the bad influence from label noise without the pre-trained backbone.

Table 1: Evaluation of EAC on noisy FER datasets.

| Method | Pretrain | 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 |
| Baseline | × | 10 | 62.55 | 77.87 | 44.73 |
| SCN (CVPR20) | × | 10 | 70.25 | 78.83 | 46.29 |
| RUL (NeurIPS21) | × | 10 | 70.75 | 80.62 | 48.53 |
| EAC (Ours) | × | 10 | 73.36 | 82.59 | 51.43 |
| Baseline | × | 20 | 56.77 | 71.98 | 41.37 |
| SCN (CVPR20) | × | 20 | 65.95 | 74.16 | 42.46 |
| RUL (NeurIPS21) | × | 20 | 68.84 | 78.96 | 46.04 |
| EAC (Ours) | × | 20 | 71.21 | 80.55 | 48.97 |
| Baseline | × | 30 | 49.70 | 69.34 | 38.47 |
| SCN (CVPR20) | × | 30 | 62.43 | 72.71 | 40.89 |
| RUL (NeurIPS21) | × | 30 | 66.92 | 76.92 | 44.77 |
| EAC (Ours) | × | 30 | 68.20 | 78.07 | 47.60 |

## 2 More results of the attention maps

We display more visualization results of the learned attention maps in Figure 1. SCN might remember noisy samples on the original images, while it can get correct predictions on some of the flipped counterparts shown in the first row of Figure 1. In the third row, we display some samples that SCN gets wrong predictions on both the original images and their flipped counterparts while our EAC can still get correct predictions.

## 3 The implementation details of experiments on CIFAR100 and Tiny-ImageNet

To show the generalization of our proposed EAC, we carry out experiments on CIFAR100 and Tiny-ImageNet. Some noisy label facial expression recognition methods are not suitable for classification tasks with a large number of classes. For example, DMUE [3] needs to train a multi-branch model to mine the latent truth of the given samples. However, when it comes to classification tasks like CIFAR-100 and Tiny-ImageNet, it is unaffordable to train a model with 101 branches to mine the latent truth, which means these methods have bad generalization ability.

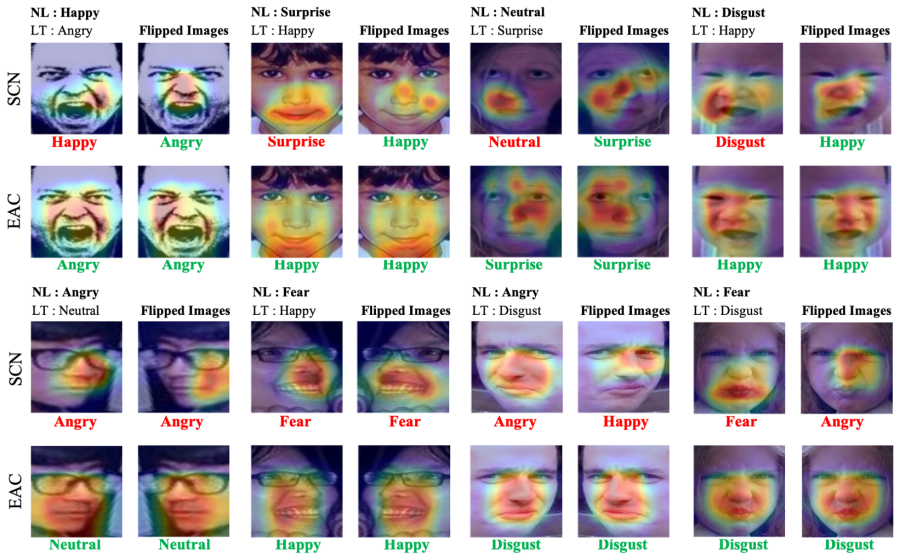



Fig. 1: The learned attention maps of SCN and EAC. EAC eliminates the negative effect of noisy labels through learning consistent attention maps. NL represents the noisy label, LT represents the latent truth. The prediction results are shown under the images.

We provide the details of the implementation in this section. We compare our proposed EAC with the baseline and SCN [4]. All the methods use ResNet-18 [2] without pre-training as the backbone network. The input images are randomly cropped to 32×32 pixels. The baseline and SCN horizontally flip the input images with a probability of 0.5 to make a fair comparison with EAC. All the methods randomly erase the input images with the probability of 0.5, the erasing scale of 0.02 to 0.25 and the erasing ratio of 0.3 to 3.3. During training, the batch size is 512. We use SGD with momentum of 0.9 and weight decay of 0.0005. The initial learning rate is 0.1. We use the Cosine Annealing learning rate scheduler with the maximum number of iterations as 200, the number of the training epochs.

## 4 Ablation Study

**Evaluation of the consistency loss weight** *lambda***.** To show the influence of $\lambda$ on the performance, we evaluate it from 0.1 to 10.0 on RAF-DB with different levels of label noise. The results are plotted in Figure 2. It is shown that the classification accuracy first increases along the $\lambda$ and then starts to decrease, which is intuitive as we use the consistency loss to regularize the classification loss. If $\lambda$ is too small, the model will overfit the noisy samples. The model degrades to the baseline method when $\lambda$ equals 0. However, when $\lambda$ is too large, the consistency loss will outweigh the classification loss. The change of the classification loss will have little effect on the model training, which leads to underfitting. The best value of $\lambda$ is 3 under 10% and 20% noise and 5 under 30% noise, which implies that we need to regularize the classification loss more when there are more noisy samples. To show the effectiveness of EAC under all circumstances, we simply set $\lambda$ as 5 in all our experiments.

**Evaluation of the erasing probability** $p$**.** We experiment with different values of erasing probability $p$ under different levels of label noise. It is shown that we need to erase more images to prevent the model from overfitting the attention maps before and after the flip when the noise ratio increases, which conforms to the choice of *lambda*, which needs to be larger to deal with more label noise.

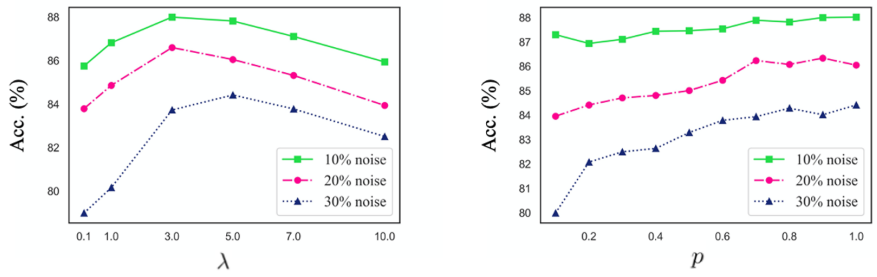



Fig. 2: Ablation study of the consistency loss weight $\lambda$ and the erasing probability $p$

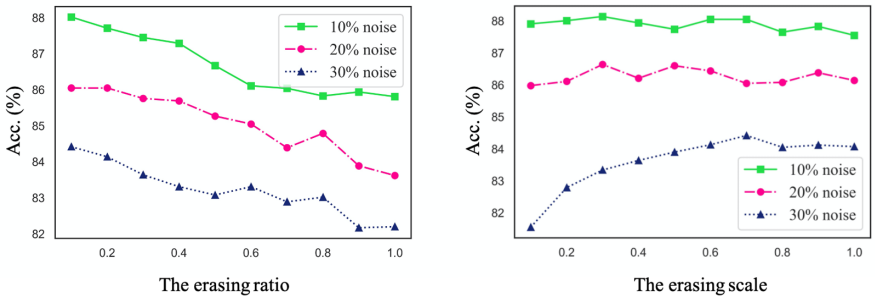



Fig. 3: Ablation study of the erasing ratio and the erasing scale.

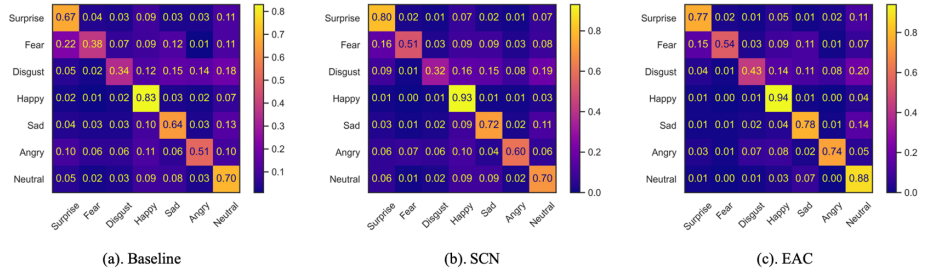
It is shown that we can choose $p$ from 0.7 to 1.0 randomly as the performance changes little.

**Evaluation of the erasing ratio.** The erasing ratio is a range of aspect ratio of the erased area. Following [5], we set the range as $(r_1, \frac{1}{r_1})$ and vary $r_1$ from 0.1 to 1.0 to evaluate its effect on the performance. The results in Figure 3 show that the performance decreases as $r_1$ increases because the range of aspect ratio of erased area narrows as $r_1$ increases, which degrades the regularization effect of the random erasing. Thus, we choose $r_1$ as 0.1 and set the erasing ratio as $(0.1, 10)$ in all our experiments.

**Evaluation of the erasing scale.** The erasing scale is a range of the proportion of erased area against the input image. Following [5], we set the lower bound of the erasing scale as a fixed value and evaluate the upper bound of the erasing scale. In this paper, we set the lower bound of the erasing scale as 0.1 and vary the upper bound from 0.1 to 1.0 to evaluate its effect on the performance. The results show that the performance is only slightly affected by the erasing scale when the noise rate is 10% or 20%. When the noise rate reaches 30%, the performance increases along with the upper bound of the erasing rate, which is intuitive as we need to erase a larger part of input images to prevent the model from remembering the noise labels. We choose the upper bound as 0.7 and set the erasing scale as $(0.1, 0.7)$ to get good performances under all circumstances.

## 5 The confusion matrices of different methods on RAF-DB

To show the improvement of EAC on different classes, we display the confusion matrices of different methods on the testset of RAF-DB after training with 30% of label noise in Figure 4. Compared with the baseline method, EAC improves the classification performance of each class. EAC improves SCN on six expression classes except for the Surprise class. The results show that EAC improves noisy label FER methods on different classes instead of a specific class with many samples.




(a). Baseline (b). SCN (c). EAC

Fig. 4: The confusion matrices of different methods on the testset of RAF-DB. EAC improves noisy label FER methods on different classes instead of a specific class with many samples.

## 6 Comparison with Attention Consistency

Attention Consistency [1] equals EAC without the imbalanced framework and the erasing module. The framework of Attention Consistency and EAC are shown in Figure 5 and Figure 6 respectively. Under 30% label noise in RAF-DB, Attention Consistency only achieves 78.29% classification accuracy on the testset, while EAC achieves 84.42%. Though Attention Consistency considers visual attention consistency under spatial transforms, it cannot eliminate the bad influence of the noisy labels as it can remember noisy samples. The difference between Attention Consistency and the baseline method under noisy label setting is that Attention Consistency has twice input images as the baseline method. However, the increase of the training data does not equal the robustness to the noisy labels. The difference between Attention Consistency and EAC lies in the imbalanced framework and the erasing module. The imbalanced framework ensures that the noisy labels will not influence the flipped images. The erasing module makes the model unable to remember the attention maps of the flipped images to minimize the consistency loss. Combined with Attention Consistency, the three modules can utilize the consistency of the attention maps before and after the flip to prevent the model from remembering the noisy samples.

## 7 Feature visualization of SCN and RUL

We provide the feature visualization results of SCN and RUL under 30% label noise in Figure 7. Though the distance between different clusters is large, SCN overfits most of the noisy samples as the noisy samples are mixed evenly with the clean samples in Figure 7 (a). In Figure 7 (b), we plot the same feature as Figure 7 (a), but labeled with the latent truth. We draw the conclusion that the learned features of SCN are negatively affected by the noisy labels as the features that related to different latent truth are mixed together, shown in the red dotted boxes. Similarly, Figure 8 shows that the features learned by RUL can

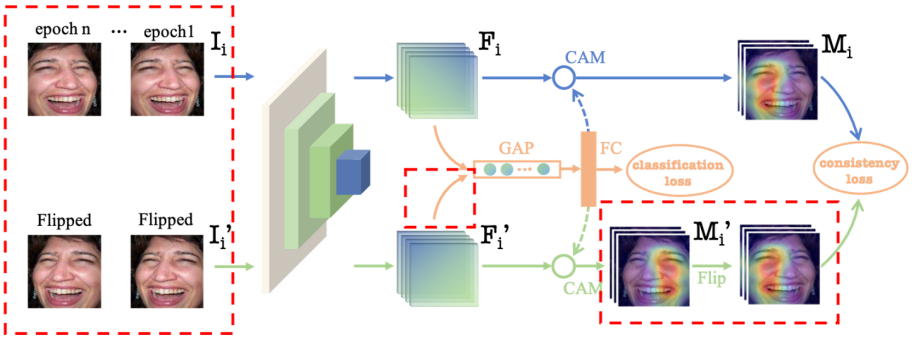



Fig. 5: The framework of Attention Consistency. The red dotted boxes mark the different parts between Attention Consistency and EAC. As the Attention Consistency does not have the imbalanced framework and the erasing part, the model might overfit the noisy labels on the flipped images, which degrades the regularization effect of the consistency loss. Thus, Attention Consistency cannot eliminate the negative effect of the noisy labels.

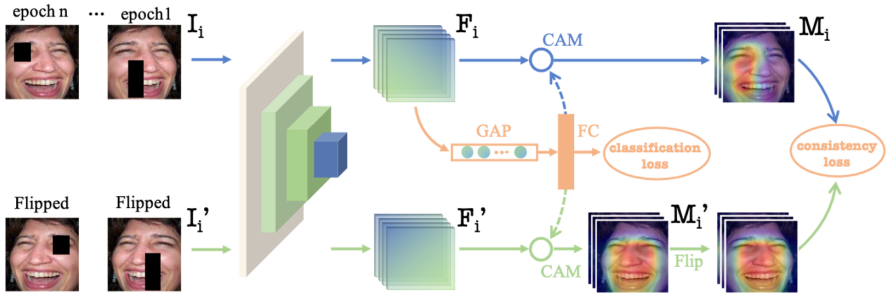
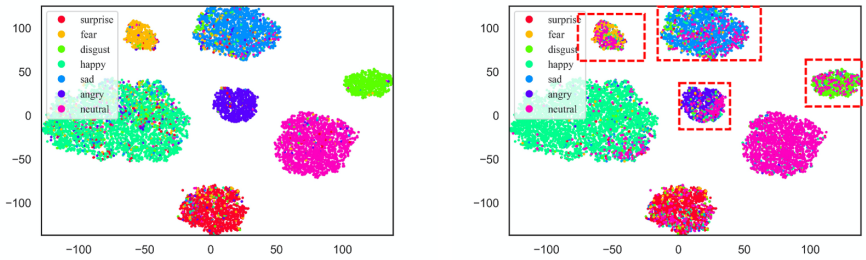


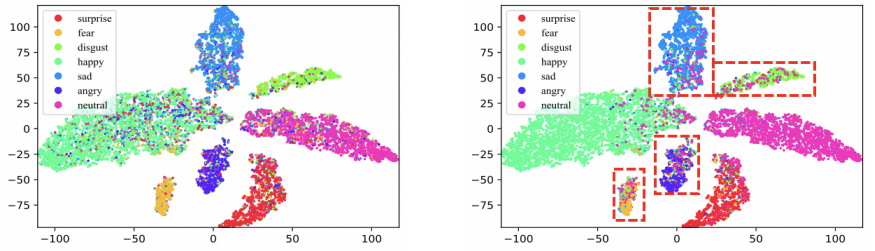

Fig. 6: The framework of our proposed EAC.

!t




(a). SCN learned features labeled with the noisy labels (b). SCN learned features labeled with the latent truth

Fig. 7: The learned features by SCN under 30% label noise. (a) is the learned features displayed with the noisy training labels. (b) is the same learned features as (a), but displayed with the latent truth. The red dotted boxes in (b) indicate the bad results of SCN as the features related to latent truth are mixed with each other.




(a). RUL learned features labeled with the noisy labels (b). RUL learned features labeled with the latent truth

Fig. 8: The learned features by RUL under 30% label noise. (a) is the learned features displayed with the noisy training labels. (b) is the same learned features as (a), but displayed with the latent truth. The red dotted boxes in (b) indicate the bad results of RUL as the features related to latent truth are mixed with each other.

not effectively suppress the noisy labels as features of noisy samples are mixed with clean samples when displayed with latent truth.

## 8 visualization of the attention maps of RUL

Figure 9 shows RUL can not acquire consistent attention on the original and flip images though we train it using both of the images before and after the flip. RUL predicts wrongly on the image circled in red, while our method predicts correctly on all these noisy images in Fig.4 of the paper.


Fig. 9: The attention maps of RUL on the original images and their flipped images.

## References

1. Guo, H., Zheng, K., Fan, X., Yu, H., Wang, S.: Visual attention consistency under image transforms for multi-label image classification. In: CVPR (2019)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
3. 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)
4. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: CVPR (2020)
5. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)