# A Traning Details

## A.1 The configuration of hyper-parameters

The training settings for ReCoT on three benchmarks are shown in Table 1. Most of the hyper-parameters are the same on the three benchmarks, except for total training epochs and epoch threshold $K$.

## A.2 Details about baseline

Since existing noisy label learning (NLL) and semi-supervised learning (SSL) methods for multi-label AU recognition are rare, we re-implement some baselines with proper modification, e.g., mean-teacher (MT [5]), noise regularization (NR [1]), bootstrapping-hard (BS-H [2]), bootstrapping-soft (BS-S [2]), Mixup [6], knowledge distillation (KD [2]). For a fair comparison, all the backbones of these methods are ResNet34 with ImageNet pretrained initialization. For MT, we change its activation function to sigmoid to perform multi-label classification. The consistency loss of MT is still mean square error (MSE), which is the same as the original paper. For NR, we just change its classification loss, and the rest remains unchanged. Same as MT, we change the activation function of BS to the sigmoid. The linear weight of noisy labels for BS-H and BS-S is set to 0.8. The difference between BS-H and BS-S is the process of predictions. BS-S uses soft prediction probability to obtain fusion labels, while BS-H uses hard prediction labels. For Mixup, we perform the same operation as BS. For KD, we set the knowledge graph of different AUs as all 1's for simplicity. The linear weight of noisy labels for KD is set to 0.5.

# B More Results

## B.1 The Effectiveness of Noisy Data

Fig.1 presents the results of supervised, SSL, and WSL methods which are all implemented by us. Compared with the supervised methods, i.e., baseline, WSL methods such as NR and KD perform better in three databases, demonstrating that noisy data could provide useful information to improve performance. Apart from EmotioNet, the overall performance of WSL approaches is better than SSL methods, indicating there exists useful information in noisy labels.

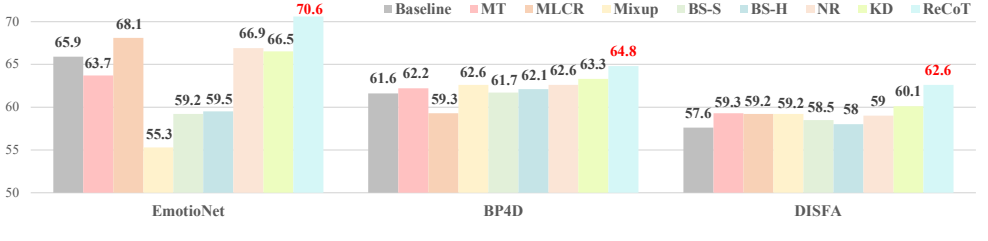
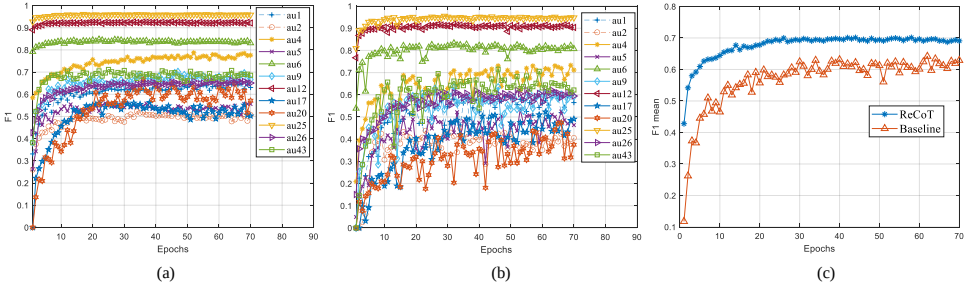



Figure 1: The comparison with baseline, MT [5], MLCR [3], Mixup [6], BS-S (H) [2], NR [1], KD [2] on three databases. Except for the results of MLCR and MT on EmotioNet, we re-implement all the other methods.

Table 1: The training configurations of hyper-parameters for ReCoT on three benchmarks.

| Config | EmotioNet | BP4D | DISFA |
|---|---|---|---|
| optimizer | Adam | Adam | Adam |
| base learning rate | 1e-3 | 1e-3 | 1e-3 |
| batch size | 100 | 100 | 100 |
| total epochs | 70 | 30 | 30 |
| $(\alpha, \beta)$ for $L_{noisy}$ | (1, 0.1) | (1, 0.1) | (1, 0.1) |
| selected ratio $\gamma$ | 0.9 | 0.9 | 0.9 |
| fusion ratio $\mu$ | 0.9 | 0.9 | 0.9 |
| momentum ratio $\phi$ | 0.9 | 0.9 | 0.9 |
| epoch threshold $K$ | 40 | 15 | 15 |
| balance coefficient $\lambda_n$ | 1 | 1 | 1 |
| balance coefficient $\lambda_{cons}$ | 100 | 100 | 100 |




(a) (b) (c)

Figure 2: F1 scores by ReCoT and ResNet-34 are reported for different testing epochs on EmotioNet without ImageNet pre-training. (a) and (b) show the F1 scores for recognizing 12 AUs by ReCoT and ResNet-34, and (c) shows the average F1 scores by two methods.

## B.2 Convergence Speed between Baseline and ReCoT

Fig. 2 shows that our ReCoT is useful for improving network convergence and the discriminative ability for AUs when exploiting the useful information from noisy data. Additionally, the training process of ReCoT is more stable compared with the baseline, since additional noisy data is utilized during the training process.

## B.3 Different Noise modeling of $net_{noisy}$

We adopt three different types of label noise modeling for $net_{noisy}$ (see Fig. (2)). Addictive label noise (Eq. 1) means modeling the label noise by $net_{noisy}$ between noisy labels and clean predictions from $net_{clean}$. Multiplicative label noise (Eq. (2)) indicates modeling the label noise transition matrix by $net_{noisy}$. Mixed label noise (Eq. (3)) denotes that the label noise consists of addictive and multiplicative label noise, which can be modeled using both the noise transition matrix and the residual connection.

$$p_{noisy} = net_{noisy}(x) + net_{clean}(x), \tag{1}$$

$$p_{noisy} = net_{noisy}(net_{clean}(x)), \tag{2}$$

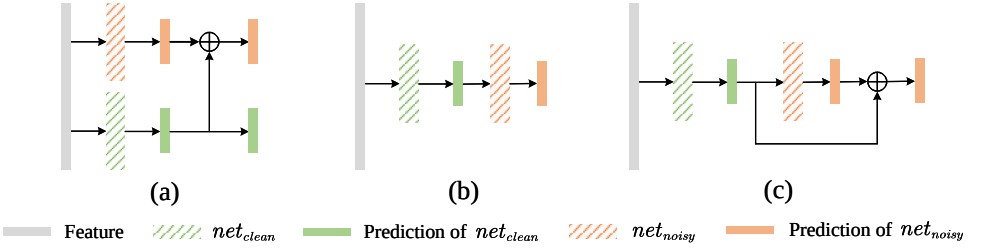



Figure 3: Three types of label noise modeling of $net_{noisy}$. (a) Addictive label noise; (b) multiplicative label noise; (c) mixed label noise.

$$p_{noisy} = net_{noisy}(net_{clean}(x)) + net_{clean}(x). \tag{3}$$

Table 2 shows the results of three types of label noise modeling. From Table 2 we can see that addictive noise modeling is good at modeling the label noise in the noisy set of EmotioNet, and thus it achieves the best result compared with the other two methods.

Table 2: Three types of label noise modeling on EmotioNet.

| **Type of label noise modeling** | **Avg.** |
|---|---|
| Addictive noise | 70.6 |
| Applicative noise | 70.2 |
| Mixed noise | 70.3 |

Table 3: The choices of encoder under pretraining on EmotioNet. **Avg.** indicates the average F1 socres of 12 AUs.

| **Type of Backbone** | **Avg.** |
|---|---|
| ResNet18 | 65.2 |
| ResNet34 | 65.9 |
| ResNet50 | 65.3 |

## B.4 The Influence of Different Encoder Choices

In the Ablation Study section of our main paper, we studied the influence when using different encoders, such as ResNet-18, ResNet-34, and ResNet-50. The detailed results are given in Table 3. We can notice that the F1 score of our method improves by 0.7% from ResNet-18 to ResNet-34, but then drops by 0.6% from ResNet-34 to ResNet-50. A possible reason is that the patterns related to AUs maybe not that complicated, and the overall modeling difficulty comes from the label confusion for the AU recognition task.

## B.5 The Influence of Different Noisy Image Scales

We briefly mentioned the influence of different noisy images Ablation Study section of our main paper. Here, we show the detailed results in Table 4. It can be observed from Table 4

Table 4: The influence of the scale of noisy images on emotioNet. **Avg.** indicates the average F1 socres of 12 AUs.

| Scale of Noisy Images | Avg. |
|---|---|
| 10,000 | 69.8 |
| 50,000 | 70.6 |
| 100,000 | 70.2 |

that the performance improves 0.8% as the number of noisy images increases from 10,000 to 50,000, but gets saturated when further increasing the scale. We assume that 50,000 images can well represent the noisy label distribution in EmotioNet.

Table 5: The results of different views on EmotioNet. **Avg.** indicates the average F1 score of 12 AUs.

| Views' Type | Avg. |
|---|---|
| View1 | 70.0 |
| View2 | 70.1 |
| The Average of Two Views | 70.6 |

## B.6 The Results of Different Views on EmotioNet

Since our ReCoT method contains two views, Table 5 shows the performance of two views and their average on EmotioNet during the inference phase. From Table 5, we can notice that the difference between two views is minor, while the average of two views is slightly better than a single view (about 0.5% higher). This suggests that when there is a low computation cost requirement, we may use a single view to perform inference, and the performance is still good; otherwise, we can use the average of two views for better performance.

# References


[1] Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. Weakly supervised image classification through noise regularization. In *Proc. CVPR*, pages 11517–11525, 2019.

[2] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proc. CVPR*, pages 1910–1918, 2017.

[3] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. In *Proc. NeurIPS*, pages 909–919, 2019.

[4] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *Proc. ICLRW*, 2014.

[5] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, volume 30, 2017.

[6] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018.