ReCoT: Regularized Co-Training for Facial Action Unit Recognition with Noisy Labels

Yifan Li[1,2]
liyifan20g@ict.ac.cn
Hu Han[1,2]
hanhu@ict.ac.cn
Shiguang Shan[1,2]
sgshan@ict.ac.cn
Zhilong Ji[3]
jizhilong@tal.com
Jinfeng Bai[3]
jfbai.bit@gmail.com
Xilin Chen[1,2]
xlchen@ict.ac.cn

[1] Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS
Beijing, China

[2] University of the Chinese Academy of Sciences
Beijing, China

[3] Tomorrow Advancing Life
Beijing, China

Abstract

Facial action unit (AU) recognition is essential for recognizing fine-grained changes in facial expression, while the demand for a large amount of accurately labeled AU data for training purposes has resulted in high labor costs. Nevertheless, massive face images are widely available and inaccurate labels can be easily obtained, especially as large vision-language pre-training models progress. This paper introduces the Regularized Co-Training (ReCoT) method, which leverages the useful information from both accurately labeled (clean) and inaccurately labeled (noisy) face images to achieve robust AU recognition. ReCoT uses a two-head network in each view, with one for clean data modeling (clean net) and the other for noisy data modeling (noisy net) by learning label noise w.r.t. the clean predictions. Additionally, a selective balanced loss is proposed for the noisy net to learn from noisy labels and alleviate the imbalanced issue in the clean net. Extensive experiments on several AU databases, including EmotioNet, BP4D and DISFA, show that ReCoT effectively leverages noisy AU data to improve the model performance. The code is available: https://github.com/JackYFL/ReCoT_BMVC2023.

1 Introduction

Facial expressions offer valuable cues about affective states, mental health and personality. They can be encoded as diverse combinations of comprehensive facial action units (AUs) according to Facial Action Coding System (FACS) [4]. Each AU encompasses a range of muscle movements, varying in intensity and position. However, deciphering the precise

[0]This research was supported in part by the National Key R&D Program of China (grant 2021ZD0111901), and the National Natural Science Foundation of China (grant 62176249).

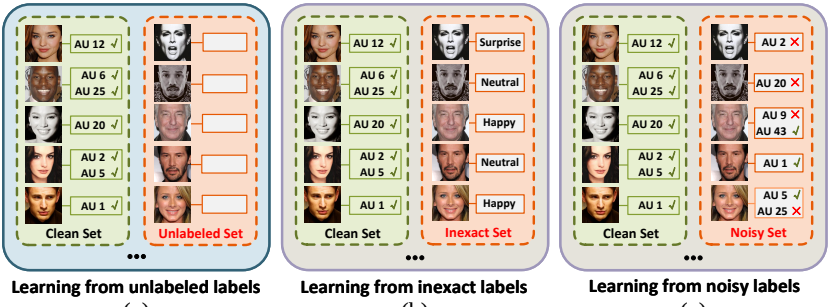

Figure 1: A comparison between SSL and WSL for AU recognition. SSL learns from a clean set with accurate AU labels and an unlabeled set without AU labels. WSL consists of two cases, i.e., (b) learning from a clean set and an inexact set with inexactly labeled AUs (e.g., emotions), and (c) learning from a clean set and a set with inaccurate (noisy) AU labels.

muscle movements corresponding to each AU can be challenging due to their subtle and ambiguous nature. Consequently, extensive training is required for a qualified AU annotation expert. Moreover, even for trained experts, manually labeling AUs is a time-consuming task. By contrast, inaccurate AU labels can be easily obtained at a low cost using pre-trained AU recognition models or employing non-professional annotators through crowd-sourcing.

While tremendous progress has been made on facial AU recognition, the limitation of accurately labeled AU data has hindered the research of AU recognition methods. Many existing approaches mainly require accurately labeled AU data [22, 28, 33, 43, 44]; thus, they can easily get over-fitting when the accurately labeled AU data is limited. Some recent works utilized unlabeled data through semi-supervised learning (SSL) to improve the AU recognition accuracy [21, 24]. While SSL-based AU recognition methods show better robustness than fully-supervised learning methods under small labeled data scenarios, as shown in Fig. 1, they cannot exploit the useful information from inexact or inaccurate AU labels (weakly supervised learning, WSL [6]) that can be easily obtained at a low cost. For example, EmotioNet [5] contains 25,000 face images with accurate AU labels and 975,000 face images with inaccurate AU labels. Facial AU recognition methods that can learn from noisy labels (LNL, a.k.a, noisy label learning) can greatly extend the application scope in practice.

LNL-based methods have achieved great success in general image classification, which can be divided into three trends, i.e., regularization [11, 18, 19, 26, 41], sample selection [7, 40] and label correction [8, 26, 32]. Regularization-based methods try to improve the model robustness by designing more robust loss functions [19, 26] or augmentation strategies [41]. However, the performance of regularization-based methods is sensitive to the degree of label noises (see Table 1). Sample selection-based methods aim to select reliable samples from noisy data via losses [7, 40] or feature similarity [14]; however, the unselected data are ignored during model training, which may affect the generalization abilities. Label correction-based methods re-annotate samples by either soft pseudo labels [32] or hard pseudo labels [16], which is easy to induce confirmation bias. In addition, most LNL-based methods only consider the single-label situation. When the clean set and noisy set contain multi-labels, there are only a limited number of approaches, such as noise regularization (NR) [11] and label cleaning network (LCN) [35].

The main contributions of this work are as follows. (1) We propose a novel regularized co-training method (ReCoT) for AU recognition with noisy labels, which uses a two-head

network (clean net and noisy net) in each of the two views to model the clean data and the label noise between clean and noisy data, respectively. Such a model exploits useful information from noisy data to improve the robustness of the clean net while reducing the risk of over-fitting. We empirically show that noisy data can offer useful information to improve performance. (2) We propose a novel selective balanced loss for the noisy net, which selects a portion of small AU negative logarithm likelihood (NLL) to update the parameters and decouples positive and negative components with individual weights. Such a loss can learn from noisy labels via noisy net and alleviate the imbalanced problem of the clean net.

2 Related Work

Existing methods using additional images and manually labeled images for AU recognition can be categorized into two folds: SSL- and WSL-based methods (see Fig. 1). SSL-based methods aims at leveraging unlabeled images to improve the model's generalization ability. WSL-based methods focus on exploiting useful information from images with inexact or inaccurate annotations. Inexact annotations can be coarse-grained labels w.r.t. the task, i.e., expression labels for AU recognition tasks. Inaccurate annotations, i.e., noisy labels, tend to contain wrong annotations and LNL-based methods are aimed for this task.

Semi-Supervised AU Recognition: To leverage the unlabeled images, self-training was proposed for AU recognition in [36, 37], which first train a teacher net on a small subset of clean data and use it to annotate unlabeled images and obtain pseudo-AU labels. Then, they jointly use the clean labels and pseudo labels with high confidence to train the whole network. However, such models may induce confirmation bias from pseudo labels, and the training time is heavy. Niu *et al.* [21] proposed a co-training method with co-regularization to leverage both labeled and unlabeled face images for semi-supervised AU recognition, while this method cannot use the effective information in noisy labels.

Weakly-Supervised AU Recognition: Zhao *et al.* [45] utilized spectral clustering to learn an embedding space for re-annotating AU labels of noisy images. However, spectral clustering may not work well when the number of images is large, and the data distribution is severely imbalanced. Moreover, this method doesn't utilize the clean set information during re-annotating which may cause bias. Peng *et al.* [23] proposed to learn from domain knowledge and expression-annotated facial images through adversarial training. They treat expression labels rather than noisy AU labels as weak labels, while the acquiring of expression labels may need another expression-based dataset, and there also exist noises in expression labels. Cui *et al.* [3] utilized Bayesian Network (BN) to capture the generic knowledge on relationships between AUs and expressions, which is then embedded into a deep learning network. Again, their method also requires expression labels as weak labels.

3 Proposed Method

3.1 Formulation

Let $\mathcal{D} = \mathcal{N} \bigcup \mathcal{C} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a facial AU dataset with L AUs in total, in which $N = N_{noisy} + N_{clean}$, $\mathcal{N} = \{(x_j^{noisy}, y_j^{noisy})\}_{j=1}^{N_{noisy}}$ is the noisy set with inaccurate AU labels, and $\mathcal{C} = \{(x_j^{clean}, y_j^{clean})\}_{j=1}^{N_{clean}}$ is the clean set with accurate AU labels. $y_i \in \mathbb{Z}^L$ denotes the AU labels of L dimensions, and $y_{i,k}$ indicates each entry of y_i. In the cases when some AUs are not visible, the AUs can be labeled as 0; otherwise, the AUs are labeled as either 1 or -1 denoting activation or not, i.e., $y_{i,k} \in \{1, -1\}$ or $y_{i,k} \in \{1, 0, -1\}$.

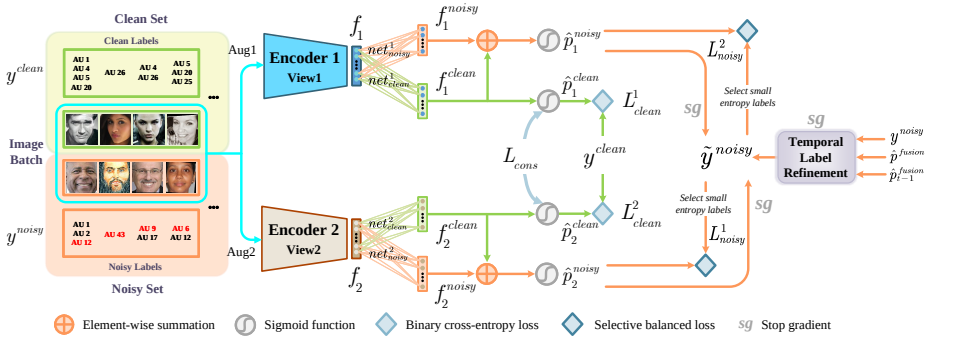

Figure 2: ReCoT uses a network with two heads (net_{clean} and net_{noisy}) per view (formed by two independent augmentations). net_{clean} performs AU classification using only the clean data. net_{noisy} aims at noisy data modeling by learning label noise which can be combined with the output feature by net_{clean} to obtain the feature for predicting the noisy AU labels. As a result, net_{noisy} works as a regularization term to avoid over-fitting of net_{clean}. Moreover, temporal label refinement is proposed to obtain more reliable labels for the noisy data, which are used to co-teach the net_{noisy} to improve the performance further.

The goal of our approach is to learn an AU classifier $\mathcal{F}$ by using both $\mathcal{N}$ and $\mathcal{C}$, which is expected to work better than an AU classifier learned from $\mathcal{C}$ alone. This is a different problem than SSL, which can only use unlabeled data instead of noisy data. In this work, we learn $\mathcal{F}$ by proposing a Regularized Co-Training approach (ReCoT), which can effectively exploit the useful information in both the clean set $\mathcal{C}$ and the noisy set $\mathcal{N}$. The overall framework of ReCoT is shown in Fig. 2, which consists of two components: regularized co-training and temporal label refinement.

3.2 Regularized Co-Training

As shown in Fig. 2, ReCoT avoids getting biased by co-training learning from two views formed by two independent augmentations. In each view, net_{noisy} is utilized to dig useful information from noisy labels and works as a regularization term to prevent net_{clean} from over-fitting to the small clean set. The overall loss L can be denoted as

$$L = \frac{1}{2}\sum_{v=1}^{2}(L^{v}_{clean} + \lambda_n L^{v}_{noisy}) + \lambda_{cons} L_{cons}, \tag{1}$$

where L^{v}_{clean}, L^{v}_{noisy}, L_{cons} indicate clean net loss, noisy net loss and consistency loss. λ_n and λ_{cons} are hyper-parameters balancing different losses. Then we'll detail each loss below.

As shown in Fig. 2, ReCoT uses a two-view learning network. For each view, we design two heads: clean net (net_{clean}) and noisy net (net_{noisy}). net_{clean} performs AU classification only using the randomly sampled face images from clean set $\mathcal{C}$. The prediction probability by net^{v}_{clean} in the v-th view can be expressed as $\hat{p}^{clean}_{v} = \sigma(f^{clean}_{v})$, where $f^{clean}_{v} = net^{v}_{clean}(f_v)$ is the features produced by clean net. Then, we can use a binary cross-entropy loss to optimize f^{clean}_{v}. Since the AU distribution is extremely imbalanced in practice, a selective learning strategy [9] is employed. The loss L^{v}_{clean} for net_{clean} in the v-th view can be denoted as

$$L^{v}_{clean} = \frac{1}{L}\sum_{k=1}^{L}\alpha^{clean}_{k}[y^{clean}_{k}\log\hat{p}^{clean}_{v,k} + (1-y^{clean}_{k})\log(1-\hat{p}^{clean}_{v,k})], \tag{2}$$

where α_k^{clean} is a selecting parameter for deciding whether this AU is used or not. For a face image with an AU label of '0', we directly set α_k^{clean} to zero to ignore its influence on loss computation. We turn the AU label '-1' to '0' to calculate the cross-entropy loss.

Different from the clean label supervised learning in net_{clean}, net_{noisy} learns label noise which can be combined with the output by net_{clean} to predict the noisy AU labels. Let $f_v^{noisy} = net_{noisy}^v(f_v)$ denote the features produced by noisy head. Then, the AU prediction probability by net_{noisy}^v in the v-th view can be expressed as

$$\hat{p}_v^{noisy} = \sigma(f_v^{clean} \oplus f_v^{noisy}), \tag{3}$$

where $\oplus$ indicates element-wise feature summation.

In order to reduce the contamination and the imbalanced problem of noisy AU labels, we propose ***selective balanced loss*** L_{noisy}^v for net_{noisy} in the v-th view, which consists of two negative logarithm likelihood (NLL): the positive NLL, and the negative one:

$$L_{noisy}^v = \frac{-\alpha}{|\mathcal{D}_p^{v'}|} \sum_{(i,k)\in\mathcal{D}_p^{v'}} \log p_{v,i,k}^{noisy} + \frac{-\beta}{|\mathcal{D}_n^{v'}|} \sum_{(i,k)\in\mathcal{D}_n^{v'}} \log(1 - p_{v,i,k}^{noisy}), \tag{4}$$

where $\alpha, \beta \in [0,1]$ are the balanced coefficients that control the weight of positive NLL and negative NLL, respectively. (i,k) means the k-th AU of the i-th face image. $\mathcal{D}_p^{v'}$ and $\mathcal{D}_n^{v'}$ indicate the selected positive and negative small AU NLL sets of the other view (v') using refined AU labels $\tilde{y}^{noisy}$, respectively. Specifically, $\left|\mathcal{D}_p^{v'}\right| = \gamma N_p$, $\left|\mathcal{D}_n^{v'}\right| = \gamma N_n$, where γ is the selected ratio, N_p and N_n indicate the numbers of positive and negative AU entries of refined noisy AU labels $\tilde{y}^{noisy}$, respectively.

Most of the existing LNL-based methods have explicit assumptions about the noisy label distributions, and their performance may drop significantly when the assumptions do not hold. While in our approach, we do not require such assumptions. As shown in Fig. 2, net_{noisy} learns label noise between the predictions for noisy and clean labels, which acts as a regularization term. As a result, the modeling capability and robustness of net_{clean} in each view can be significantly improved compared to learning from only clean data. Additionally, L_{noisy} can exploit the useful information in noisy labels by selecting reliable ones and co-teaching net_{noisy}, and it can also alleviate the imbalanced issue of net_{clean} by decoupling positive and negative AU NLL with two different weights α, β.

In addition to the regularization in each view, ReCoT also uses a cross-view consistency loss L_{cons} to assure that the AU predictions for both the clean and noisy data by the net_{clean} heads in two views should be as consistent with each other [21, 25]

$$L_{cons} = \frac{1}{L}\sum_{k=1}^{L}[H(\frac{\hat{p}_{1,k}^{clean} + \hat{p}_{2,k}^{clean}}{2}) - \frac{H(\hat{p}_{1,k}^{clean}) + H(\hat{p}_{2,k}^{clean})}{2}], \tag{5}$$

where $H(p(x)) = -(p(x)\log p(x) + (1-p(x))\log(1-p(x)))$ denoting the entropy of AU x.

3.3 Temporal Label Refinement

Assume that the fusion prediction probability $\hat{p}^{fusion}$ can be expressed as:

$$\hat{p}^{fusion} = \mu\,\bar{p}^{noisy}(x_{noisy}) + (1-\mu)\bar{p}^{clean}(x_{noisy}), \tag{6}$$

Table 1: F1 score (in%) for recognition of 12 AUs by different methods on the EmotioNet database. Baseline is ImageNet pre-trained ResNet-34 on $\mathcal{C}$. Except for the results with *, all the other results are taken directly from paper [21].

Method/AU		1	2	4	5	6	9	12	17	20	25	26	43	Avg.
SL	Baseline*	60.2	45.4	72.8	51.5	81.4	65.4	91.2	53.0	46.7	95.0	61.7	66.6	65.9
SSL	MLCT [63]	57.8	44.8	[73.7]	50.1	82.8	58.1	[91.8]	44.8	37.1	[95.1]	[61.6]	63.4	63.4
	MT [24]	55.5	46.3	71.7	48.6	81.6	61.7	91.0	46.7	43.5	94.7	60.2	63.9	63.7
	CoT [1]	[58.3]	[48.4]	70.0	[50.4]	[83.1]	[64.4]	91.7	[49.9]	[47.1]	95.0	60.0	[66.9]	[65.5]
	MLCR [21]	**61.4**	**49.3**	**75.9**	**54.1**	**83.5**	**68.3**	**92.0**	**50.8**	**53.5**	**95.2**	**65.1**	**68.1**	**68.1**
WSL	Mixup* [10]	40.9	29.3	65.1	39.0	77.6	54.9	90.8	32.8	41.9	93.4	46.1	52.3	55.3
	BS-S* [23]	49.0	36.4	67.8	45.6	78.9	54.5	90.6	37.8	46.0	93.4	50.2	59.9	59.2
	BS-H* [23]	48.0	36.5	67.9	45.9	79.1	55.5	90.6	37.7	49.6	93.4	50.3	59.0	59.5
	KD* [13]	59.4	45.0	[74.0]	52.8	[82.1]	[66.1]	[91.7]	50.8	53.2	[95.1]	59.6	[68.2]	66.5
	NR* [21]	[61.7]	[49.5]	73.0	[53.1]	81.3	65.0	91.3	[51.9]	[54.3]	94.8	[61.0]	66.2	[66.9]
	ReCoT*	**64.2**	**52.0**	**77.8**	**55.6**	**83.9**	**70.5**	**92.5**	**58.7**	**57.3**	**95.9**	**65.8**	**72.9**	**70.6**

where μ is the fusion ratio, $\bar{p}^{noisy}(x_{noisy})$ and $\bar{p}^{clean}(x_{noisy})$ indicate the average probability of net_{noisy} and net_{clean} from two views to noisy images x_{noisy}: $\bar{p}^{clean}(x_{noisy}) = \frac{1}{2}\sum_{v=1}^{2}\hat{p}_v^{clean}(x_{noisy})$, $\bar{p}^{noisy}(x_{noisy}) = \frac{1}{2}\sum_{v=1}^{2}\hat{p}_v^{noisy}(x_{noisy})$. In order to fuse the information in the previous epochs, we propose to adopt temporal filtering by performing exponential smoothing on $\hat{p}^{fusion}$:

$$\hat{p}_t^{fusion} = \phi\,\hat{p}_{t-1}^{fusion} + (1-\phi)\hat{p}^{fusion}, \tag{7}$$

where ϕ is the momentum ratio, $\hat{p}_{t-1}^{fusion}$ is the fusion prediction probability of the last epoch. Then, the refined labels y_t^{fusion} is expressed as $y_t^{fusion} = \mathbb{1}[\hat{p}_t^{fusion} > 0.5]$. Since the modeling capacity of net_{noisy} is not strong enough at the beginning of the training, we regard noisy labels as the refined noisy labels $\tilde{y}_t^{noisy}$ before the network is converged at the K-th epoch.

We then use the refined AU labels $\tilde{y}^{noisy}$ to retrain the net_{noisy} in both views. Note that the gradient of $\overline{p}^{noisy}$ and $\overline{p}^{clean}$ has been stopped when computing $\tilde{y}^{noisy}$. During the inference phase, we use the average AU prediction of the two net_{clean}^{v} in two views as our final AU recognition results. Our rationale is that when the co-training network gradually converges, the refined labels $\tilde{y}^{noisy}$, which fuses the temporal information of net_{clean} and net_{noisy}, becomes more reliable than the original noisy labels y^{noisy}.

4 Experiment

We provide evaluations for the proposed approach on three widely used facial AU databases (EmotioNet [5], BP4D [42] and DISFA [20]). All the datasets are publicly available, so there are no Institutional Review Board issues.

4.1 Databases and Protocols

EmotioNet is an in-the-wild AU database containing 975,000 images with 12 pseudo-AU labels ({-1, 1}, noisy labels) produced by the algorithm in [5] and 25,000 images with 23 manual AU labels ({-1, 0, 1}, clean labels). Following the protocol in [21], we also use 12 AUs (1, 2, 4, 5, 6, 9, 12, 17, 20, 25, 26, 43) for evaluations. Since some download links provided with the data are broken, we are only able to get 20,722 face images with manual

Table 2: F1 score (in%) for recognition of 12 AUs by ReCoT on the BP4D database. Baseline is ImageNet pre-trained ResNet-34 on $\mathcal{C}$. Except for the results with *, all the other results are taken directly from the original papers.

Method/AU		1	2	4	6	7	10	12	14	15	17	23	24	Avg.
SL	DRML [46]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
	EAC-Net [15]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
	DSIN [6]	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
	CMS [7]	49.1	44.1	50.3	79.2	74.7	80.9	**88.3**	63.9	44.4	60.3	41.4	51.2	60.6
	LP-Net [35]	43.4	38.0	54.2	77.1	76.7	[83.8]	87.2	63.3	45.3	60.5	[48.1]	54.2	61.0
	ARL [29]	45.8	39.8	55.1	75.7	77.2	82.3	86.6	58.8	47.6	62.1	47.4	[55.4]	61.1
	JÂA-Net* [31]	47.2	41.6	49.1	77.2	77.5	82.9	85.8	63.4	50.8	62.5	47.2	52.7	61.5
	Baseline*	49.3	43.6	57.2	77.6	[78.4]	83.0	85.7	60.2	49.5	61.6	47.0	47.0	61.6
	SRERL [13]	46.9	45.3	55.6	77.1	[78.4]	83.5	[87.6]	60.6	[52.2]	**63.9**	47.1	53.3	62.9
	HMP-PS [33]	[53.1]	46.1	56.0	76.5	76.9	82.1	86.4	[64.8]	51.5	63.0	**49.9**	54.5	63.4
	SEV-Net [34]	**58.2**	**50.4**	[58.3]	**81.9**	73.9	**87.8**	87.5	61.6	**52.6**	62.2	44.6	47.6	[63.9]
	ATCM [12]	51.7	[49.3]	**61.0**	[77.8]	**79.5**	82.9	86.3	**67.6**	51.9	[63.0]	43.7	**56.3**	**64.2**
SSL	MLCR* [20]	[45.5]	[40.0]	[48.6]	[76.4]	[77.0]	[81.6]	[85.7]	**62.8**	[39.2]	[61.8]	[41.0]	**51.8**	[59.3]
	MT* [32]	**50.2**	**42.5**	**57.9**	**77.1**	**79.5**	**83.9**	**86.7**	[60.6]	**51.3**	**62.2**	**47.9**	[47.0]	**62.2**
WSL	BS-S* [26]	51.2	42.5	55.7	76.5	78.8	83.4	87.0	59.1	48.8	62.5	45.7	49.4	61.7
	BS-H* [26]	52.1	43.6	55.3	77.7	78.8	83.9	86.5	59.3	50.5	62.8	45.8	49.2	62.1
	Mixup* [40]	**54.6**	43.7	57.1	**79.5**	79.3	[84.6]	[87.7]	**62.8**	[53.2]	[63.6]	46.3	39.5	62.6
	NR* [41]	50.1	44.1	56.0	77.4	78.9	83.5	86.6	59.1	50.0	62.3	[49.7]	[53.7]	62.6
	KD* [18]	[52.2]	[44.7]	**59.7**	78.3	[79.8]	84.4	86.8	61.0	50.6	62.1	48.4	51.2	[63.3]
	ReCoT*	51.5	**47.8**	[58.9]	[79.2]	**80.2**	**84.9**	**88.4**	[61.6]	**53.3**	**64.6**	**51.8**	**55.4**	**64.8**

annotations, in which we use 15, 000 randomly selected face images for training, and the remaining 5, 722 images for testing. We perform a random dataset split three times to run the experiments to avoid biased results. **BP4D** contains 328 videos from 41 subjects including 23 females and 18 males. There are 12 AUs (1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, 24) and about 140,000 frames with AU labels ({-1, 1}). We follow [17, 28], and use a subject-exclusive 3-fold cross-validation testing protocol. **DISFA** includes 27 videos recorded from 12 females and 15 males, and 8 of the 12 AUs are used for evaluation. There are about 130,000 frames, with each frame annotated with an AU intensity from 0 to 5. Following [21, 44], an AU with intensity equal to or greater than 2 is considered to be activated; otherwise, it is considered inactivated ({-1, 1}).

We use dlib to detect five facial landmarks and use them to align and crop the face images into 240×240. During training, each face image is randomly cropped to 224×224, with random horizontal flipping, grayscale, and color jittor for data augmentation. During testing, we only use center cropping. Following [21], we also use 50,000 face images with pseudo labels in EmotioNet as our noisy dataset for all three databases. For BP4D and DISFA, we generate their noisy labels by using a baseline trained on BP4D and DISFA with according fold, respectively.

In this paper, "clean set" refers to the subset that has the manually verified labels, while "noisy set" refers to the remaining training data with inaccurate labels (pseudo-labels in our setting). However, we don't impose a specific limitation on the source of label noise.

4.2 Training Details and Evaluation Metrics

Training details. We use two ResNet-34 [10] as encoders. net_{clean} and net_{noisy} are two-layer perceptron structures ($512 \times L$). We use Adam optimizer [13] with a constant learning rate of 0.001 to optimize the whole network. We use ImageNet pre-trained ResNet-34 as

Table 3: F1 score (in%) results for recognition of 8 AUs on DISFA. Baseline is ImageNet pre-trained ResNet-34 on $\mathcal{C}$. Except for the results with *, all the other results are taken directly from the original papers.

Method/AU		1	2	4	6	9	12	25	26	Avg.
SL	DRML [42]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
	EAC-Net [21]	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
	DSIN [9]	42.4	39.0	68.1	28.6	46.8	70.8	90.4	42.2	53.6
	SRERL [15]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
	LP-Net [27]	29.9	24.7	[72.7]	46.8	49.6	72.9	93.8	65.0	56.9
	CMS [24]	40.2	44.3	53.2	**57.1**	50.3	73.5	81.1	59.7	57.4
	Baseline*	45.5	36.0	68.0	40.1	40.8	74.2	94.2	61.0	57.6
	ARL [29]	43.9	42.1	63.6	41.8	40.0	76.2	[95.2]	66.8	58.7
	JÂA-Net* [30]	38.6	41.7	68.0	41.2	39.5	[77.0]	94.3	**69.5**	58.7
	SEV-Net [39]	**55.3**	**53.1**	61.5	53.6	38.2	71.6	**95.7**	41.5	58.8
	HMP-PS [31]	38.0	45.9	65.2	50.9	[50.8]	76.0	93.3	[67.6]	61.0
	ATCM [14]	[46.1]	[48.6]	**72.8**	[56.7]	50.0	72.1	90.8	55.4	**61.5**
SSL	MLCR* [21]	[42.7]	[31.8]	[65.6]	**47.5**	**49.9**	**77.0**	[93.7]	**65.2**	[59.2]
	MT* [34]	**48.4**	**40.0**	**66.3**	[43.9]	[43.7]	[74.6]	**93.8**	[63.2]	**59.3**
WSL	BS-H* [26]	40.4	37.7	**68.4**	42.0	45.9	74.6	94.0	61.2	58.0
	BS-S* [26]	42.9	**41.7**	67.2	41.1	48.1	74.3	94.2	58.9	58.5
	NR* [40]	43.2	35.2	[67.9]	[45.8]	48.3	75.1	94.5	[62.2]	59.0
	Mixup* [40]	[48.7]	40.6	64.9	45.3	45.8	[77.4]	93.4	57.6	59.2
	KD* [13]	48.0	38.5	69.1	45.3	[50.3]	76.0	[94.6]	59.4	[60.1]
	ReCoT*	**51.3**	36.2	66.8	**50.1**	**52.4**	**78.8**	**95.3**	**69.7**	**62.6**

encoder and finetune the whole network when testing on EmotioNet, BP4D and DISFA. The maximum training epochs for EmotioNet, BP4D, and DISFA are 70, 25, and 25, respectively. In each epoch, similar to [21], we use a batch size of 100, and randomly sample face images from clean data and noisy data. More details about hyper-parameters could be found in Appendix. All the experiments are conducted on GeForce GTX 3090 GPU using PyTorch.

Evaluation Metrics. Following [17, 22, 30], we report F1 score and the average F1 score (Avg.) on three datasets (Table 1, 2, 3), where the best and the second best results are indicated with bold and brackets, respectively.

4.3 AU Recognition Results

Results on EmotioNet. We compare ReCoT with several SOTA SSL-based AU recognition methods. For a fair comparison, these methods also use ResNet-34 as backbone and the same testing protocol. ResNet-34 using ImageNet pre-trained model and finetuned with only clean data is also used as a baseline. The experiment results are presented in Table 1, in which the result for baseline, Mixup, BS, KD and NR are implemented by us, and the results for the other SOTA methods are taken from [21]. The details about re-implement algorithms can be found in Appendix.

We can see that our method outperforms all the SOTA methods by a large margin. This indicates that although the noisy labels are inaccurate, there is useful information that can be exploited to improve the network performance. Furthermore, the results suggest leveraging noisy labels can avoid model over-fitting and improve generalization ability. In addition, since the original noisy labels in EmotioNet include more noises (Table 4), BS and Mixup perform worse than KD and NR which are more robust to noisy labels.

Results on BP4D. Table 2 reports the F1 score of the SOTA supervised learning methods, SSL method and WSL methods on BP4D. ResNet-34 trained with only clean data is used as a baseline, which is the same as EmotioNet. We run MLCR [21], JÂA-Net [30] and MT [34] using its open-sourced code.

Table 4: The ablation study of ReCoT on EmotioNet. Baseline ($\mathcal{C}$) means baseline is trained on $\mathcal{C}$. Baseline ($\mathcal{N}$) and Baseline ($\mathcal{N}+\mathcal{C}$) are defined in the same way. Baseline + $\boldsymbol{net}_{noisy}$ indicates baseline with two heads that is trained on $\mathcal{N}+\mathcal{C}$. **CoT** and **TLR** represent co-training without $\boldsymbol{net}_{noisy}$ and temporal label refinement, respectively.

Method	Avg.	
	w/o pre	w pre
Baseline ($\mathcal{C}$)	63.1	65.9
Baseline ($\mathcal{N}$)	45.2	45.6
Baseline ($\mathcal{C}+\mathcal{N}$)	56.6	59.2

(a) Different training sets

Method	Avg.	
	w/o pre	w pre
Baseline + $\boldsymbol{net}_{noisy}$	65.5	66.9
Baseline + ensemble	65.4	66.8
Baseline + **CoT**	68.6	69.5
Baseline + **CoT** + $\boldsymbol{net}_{noisy}$	69.0	69.8
Baseline + **CoT** + $\boldsymbol{net}_{noisy}$ + **TLR**	69.2	70.1

(b) Different components

Method	α	β	γ	Avg.	
				w/o pre	w pre
ReCoT	1	1	1	69.2	70.1
ReCoT	1	1	0.9	69.3	70.1
ReCoT	1	0.5	0.9	69.4	70.3
ReCoT	1	0.1	0.9	69.7	70.6

(c) Hyper-parameter of L^{v}_{noisy}

It can be observed that ReCoT overall outperforms all previous works w.r.t. the average F1 score. Compared with SSL-based methods such as MLCR and Mean-Teacher, we improve around 5.5% and 2.6% on average F1 score by learning from noisy labels. Our method also performs better than other WSL methods which shows that our method is efficient in exploiting extra information from noisy data. Compared with SL methods, such as LP-Net, JÂA-Net and ATCM, which require landmarks or attention maps, we also achieve better results by learning from noisy data.

Results on DISFA. We compare with supervised, SSL and WSL AU recognition methods which are the same as BP4D. The results are presented in Table 3.

The results show that our results perform better than SSL-based methods, which indicate that noisy labels could provide useful information via net_{noisy} to improve performance. Compared with WSL-based methods, we also perform well, which is contributed to co-training which has two views to provide diverse features. ATCM achieves a slightly higher F1-score than our method, particularly for AUs 2, 4 and 6. This is because ATCM leveraged facial-landmarks-based attention maps and a larger backbone network (Inception V3) for feature learning. Moreover, the big illumination, pose, and background gaps between DISFA and extra dataset EmotioNet also bring additional challenges for AU recognition.

4.4 Ablation Study

We provide ablation studies in Table 4 to investigate the effectiveness of individual components in ReCoT, considering both finetuning from ImageNet pre-trained model (pre) and learning from scratch (w/o pre) to better validate the effectiveness of each modules.

The effectiveness of two heads. The F1 scores of the baseline model with and without using pre-training increase by 1% and 2.4%, respectively after using two heads per view (net_{noisy} in Table 4 (b)). Similarly, the F1 scores of co-training with and without using pre-training improve by 0.3% and 0.4%, after using two heads. This suggests that our two-head structure is effective in exploiting useful information from noisy data to improve the model's robustness. Since the noises of noisy labels in EmotioNet are heavy, the performance of baseline gets worse when using $\mathcal{N}$ (only 45.6% in $\mathcal{N}$ and 59.2% in both $\mathcal{N}$ and $\mathcal{C}$). Because ReCoT uses net_{noisy} to decouple noisy data from net_{clean}, which could alleviate the damage of noisy labels and make use of noisy labels. Compared with SSL methods such as MT and MLCR, ReCoT also achieves remarkable results (see Table 1, 2, 3), demonstrating that two heads could improve performance and robustness by learning from noisy labels.

The effectiveness of co-training. After adding a co-training module, the F1 score increases 3.6% and 6.5% compared with the baseline under pre-training and no pre-training conditions. While since ensemble learning just utilizes the fusion outputs of two nets, the individual module can not promote each other and the performance is worse than co-training.

This indicates that co-training is effective in improving the performance of both two views.

The effectiveness of L_{noisy}. We can see from Table 4 (c) that the F1 score improves about 0.4% and 0.5% for ReCoT compared with the one without L_{noisy}. Moreover, since the noisy labels are imbalanced, the performance of ReCoT improves further when decreasing β, which demonstrates the effectiveness of the positive and negative NLL decoupling. Moreover, the performance has been slightly improved when selecting from noisy labels.

The effectiveness of TLR. As shown in Table 4 (b), F1 score improves about 0.3% after adding temporal label refinement, which could provide more reliable pseudo-AU labels for net_{noisy} by fusing the temporal information of both net_{clean} and net_{noisy}.

Additionally, we investigate other factors related to our work in the Appendix, such as the effectiveness of noisy data, the convergence speed between baseline and ReCoT, the influence of encoder choice, the scale of the noisy image and modeling of label noise, etc.

5 Conclusion

In this work, we propose a regularized co-training approach (ReCoT) to learn a robust AU recognition model by using both clean data with accurate AU labels and noisy data with inaccurate AU labels. ReCoT uses a two-head network in each of two views: one for clean data modeling and the other for noisy data modeling. As a result, the noisy net can work as a regularization term modeling label noise to exploit the useful information from the noisy set to improve the performance and avoid over-fitting. We also propose selective balanced loss to learn from noisy labels and reduce the imbalanced problems of a clean net. We wish ReCoT could be employed to leverage the knowledge of the vision-language pre-training models in future applications for multi-label classification tasks.

References

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. CoLT*, pages 92–100, 1998.

[2] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proc. ECCV*, pages 298–313, 2018.

[3] Zijun Cui, Tengfei Song, Yuru Wang, and Qiang Ji. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. In *Proc. NeurIPS*, volume 33, 2020.

[4] Paul Ekman and Erika L Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[5] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proc. CVPR*, pages 5562–5570, 2016.

[6] Eric Granger, Patrick Cardinal, et al. Weakly supervised learning for facial behavior analysis: A review. *arXiv preprint arXiv:2101.09858*, 2021.

[7] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proc. NeurIPS*, volume 31, 2018.

[8] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proc. ICCV*, pages 5138–5147, 2019.

[9] Emily Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *Proc. AAAI*, volume 32, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

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

[12] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proc. CVPR*, pages 7680–7689, 2021.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2014.

[14] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proc. CVPR*, pages 5447–5456, 2018.

[15] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proc. AAAI*, volume 33, pages 8594–8601, 2019.

[16] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proc. ICLR*, 2019.

[17] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE TPAMI*, 40(11):2583–2596, 2018.

[18] 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.

[19] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Proc. NeurIPS*, volume 33, pages 20331–20342, 2020.

[20] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE TAC*, 4(2): 151–160, 2013.

[21] 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.

[22] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proc. CVPR*, pages 11917–11926, 2019.

[23] Guozhu Peng and Shangfei Wang. Weakly supervised facial action unit recognition through adversarial training. In *Proc. CVPR*, pages 2188–2196, 2018.

[24] Guozhu Peng and Shangfei Wang. Dual semi-supervised learning for facial action unit recognition. In *Proc. AAAI*, volume 33, pages 8827–8834, 2019.

[25] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proc. ECCV*, pages 135–152, 2018.

[26] 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.

[27] Nishant Sankaran, Deen Dayal Mohan, Srirangaraj Setlur, Venugopal Govindaraju, and Dennis Fedorishin. Representation learning through cross-modality supervision. In *Proc. FG*, pages 1–8, 2019.

[28] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proc. ECCV*, pages 705–720, 2018.

[29] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE TAC*, 2019.

[30] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaa-net: Joint facial action unit detection and face alignment via adaptive attention. *IJCV*, 129(2):321–340, 2021.

[31] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proc. CVPR*, pages 6267–6276, 2021.

[32] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proc. CVPR*, pages 5552–5560, 2018.

[33] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proc. ICCV*, pages 12899–12908, 2021.

[34] 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.

[35] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proc. CVPR*, pages 839–847, 2017.

[36] Pengcheng Wang, Zihao Wang, Zhilong Ji, Xiao Liu, Songfan Yang, and Zhongqin Wu. Tal emotionet challenge 2020 rethinking the model chosen problem in multi-task learning. In *Proc. CVPRW*, pages 412–413, 2020.

[37] Philipp Werner, Frerk Saxen, and Ayoub Al-Hamadi. Facial action unit recognition in the wild with multi-task cnn self-training for the emotionet challenge. In *Proc. CVPRW*, pages 410–411, 2020.

[38] Yuying Xing, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Multi-label co-training. In *Proc. IJCAI*, pages 2882–2888, 2018.

[39] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proc. CVPR*, pages 10482–10491, 2021.

[40] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption. In *Proc. ICML*, pages 7164–7173, 2019.

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

[42] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. *IVC*, 32(10):692–706, 2014.

[43] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proc. CVPR*, pages 2207–2216, 2015.

[44] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proc. CVPR*, pages 3391–3399, 2016.

[45] Kaili Zhao, Wen-Sheng Chu, and Aleix M Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *Proc. CVPR*, pages 2090–2099, 2018.