Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

1-2021

Facial emotion recognition with noisy multi-task annotations

S. ZHANG

Zhiwu HUANG
Singapore Management University, zwhuang@smu.edu.sg

D.P. PAUDEL

Gool L. VAN

Citation

1

Facial Emotion Recognition with Noisy Multi-task Annotations

Siwei Zhang[1]   Zhiwu Huang[1]   Danda Pani Paudel[1]   Luc Van Gool[1,2]

[1]Computer Vision Lab, ETH Zürich, Switzerland   [2]VISICS, KU Leuven, Belgium

`siwei.zhang@inf.ethz.ch`   `{zhiwu.huang,paudel,vangool}@vision.ee.ethz.ch`

Abstract

Human emotions can be inferred from facial expressions. However, the annotations of facial expressions are often highly noisy in common emotion coding models, including categorical and dimensional ones. To reduce human labelling effort on multi-task labels, we introduce a new problem of facial emotion recognition with noisy multi-task annotations. For this new problem, we suggest a formulation from the point of joint distribution match view, which aims at learning more reliable correlations among raw facial images and multi-task labels, resulting in the reduction of noise influence. In our formulation, we exploit a new method to enable the emotion prediction and the joint distribution learning in a unified adversarial learning game. Evaluation throughout extensive experiments studies the real setups of the suggested new problem, as well as the clear superiority of the proposed method over the state-of-the-art competing methods on either the synthetic noisy labeled CIFAR-10 or practical noisy multi-task labeled RAF and AffectNet. The code is available at `https://github.com/sanweiliti/noisyFER`.

1. Introduction

As a window to the mind, face expresses various human emotions and intents in everyday life. This leads to a common assumption that a person's emotional state can be readily inferred from his or her facial movements. To automatically recognize facial expressions of emotions, large amounts of datasets (e.g., [31, 53, 63, 6, 11, 37, 28, 24, 25]) and machine learning methods (e.g., [59, 46, 38, 3, 30, 21, 65, 58, 29]) have been suggested. However, how people communicate and understand basic categorical emotions (i.e., happiness, sadness, anger, disgust, surprise and fear) vary substantially across cultures, situations, and even across people within the same situation. In addition, it is often challenging for people to distinguish several facial
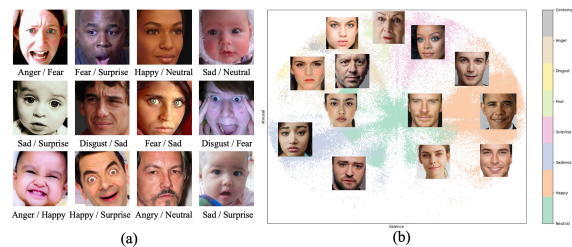

Figure 1. Illustration of (a) biased annotations on categorical emotions [37], (b) the association between categorical and dimensional emotions.

emotion pairs such as, anger vs. disgust, and surprise vs. fear. Fig. 1(a) shows some examples where two experts perceive emotions differently. Moreover, the challenge of correctly annotating emotions increases dramatically when people are asked to annotate dimensional emotions, i.e., valence and arousal values, which are typically defined within a continuous range of $[-1, 1]$. Therefore, biased annotations of facial expressions are inevitable and ubiquitous. On the other hand, as illustrated in Fig.1 (b), the categorical and dimensional labels have close correlation, despite of them being targeted for different tasks, i.e., emotion recognition vs. affect prediction.

In this paper, we explore a new problem of facial emotion recognition from noisy multi-task labels. To reduce human efforts for labelling, we suggest to make use of cheap annotations, allowing noisy labels of the kind typically obtained from the web collection or using non-expert annotators. In fact, such labels for emotion already do exist in the form of different coding models across various datasets [11, 37, 28]. Among many, two most commonly used facial emotion coding models are categorical and dimensional. Therefore, we suggest to learn emotion recognition from multi-task labels of these two kinds. Our general observation indicates that the association between image data and the available emotion labels, although noisy, are often correct when compared to the incorrect ones, for both coding models considered. Unfortunately though, directly learning from such labels results in undesired outcome during inference. In-

spired by this observations, we advice to benefit from multiple labels per image hereby posing a problem of learning from related multi-task labels of different encoding models. Our proposed formulation addresses this problem from the perspective of joint distribution match, which aims at exploiting the correlations among data and multi-task labels to reduce the impact of the labels' noise.

While the formulation facilitates us to better treat the new problem, it is challenging to achieve a unified model to play the trade-off between the basic emotion prediction and the constraint on joint distribution match. Furthermore, it is known that exact modeling of probability density function is computationally intractable [45] for all but the most trivial cases. Additionally for our case, due to the high heterogeneity between data and multi-task labels, it is non-trivial to learn their joint distribution. To cope with the second issue, we instead reduce the data distribution modeling to generative modeling of distributions, using generative adversarial networks [15], which has proved the empirical strength as generative models of arbitrary data distributions. For the third issue, we follow the idea of canonical correlation analysis [51] to learn the correlation (or joint distribution) of the heterogeneous data using multi-stream projections, with which a common space is pursued reducing the heterogeneity. The exploration of joint distribution learning enables a good collaboration between the emotion prediction and joint distribution learning tasks in an unified adversarial learning game.

In summary, this paper offers several contributions to address the practical case of facial emotion recognition. Major three of them are listed below:

- We suggest a new problem of facial emotion recognition with noisy multi-task labels, which targets for readily available cheap multi-task annotations.

- To address the proposed problem, we propose a generalized formulation with explicit joint and marginal distribution match among data and the heterogeneous multi-task labels.

- In our formulation, we introduce a new adversarial learning model to optimize the training of emotion prediction with the joint and marginal distribution based constraint, which is shown to be suitable for the newly proposed problem.

2. Related Work

Learning from noisy labels. A number of approaches have been proposed for learning noisy labels. There are several different lines to address this issue. The first line aims to parameterize the noise by a transition matrix indicating the probability to mislabel one class into another class [48, 14, 50, 39]. [39] performs a forward correction

and a backward correction to the noisy labels given a transition matrix. [48, 14] add a linear layer on top of the network to learn the transition probability jointly with the prediction network. The second line learns the label distribution [13, 49, 56]. For example, [13] introduces deep label distribution learning by assuming a predefined distribution for the true labels and minimizes the Kullback-Leiber divergence. [49, 56] optimize both the network parameters and the label predictions as the label distributions to correct noise. Some works explore this topic with two simultaneous networks [20, 33, 16, 57, 32, 27] updated iteratively. There are other methods such as sample weighting strategy to measure the confidence level of each sample [20, 61], or robust loss functions [55, 40, 62] from the loss level. Learning with multiple noisy labels has also been explored by estimating ground truth with crowdsourcing [66] or soft labels [18]. For emotion recognition with inconsistent labels, [58] extents the single noise learning formulation by incorporating a separate transition matrix for each annotator. [50] proposes to add a regularization term to minimize the trace of the transition matrix for the convergence to the true label.

To summarize, previous works generally tackle the noise label issue by transition matrix, label distribution or other classification-specific methods, therefore are incapable of dealing with multi-task involving both discrete and continuous labels, while ignoring the joint distribution between the image and labels which potentially contains richer information to benefit learning.

Generative modeling for joint distribution learning. Following generative adversarial network [15], there are emerging a few works that explore joint distribution learning in unsupervised inference tasks or conditional generation [10, 7, 19, 8, 42, 4, 12]. For instance, [10, 7, 2] approximate the posterior distribution by matching two joint distributions from an inference network and a generation network in an unsupervised or semi-supervised manner. With a slight difference, [4] distinguishes three joint distributions including the image-label pair from data, while [12] incorporates two discriminators to distinguish among encoder-decoder distributions and fake-true distributions. More recently, [42] focuses on image synthesis from both marginal distributions and conditional distributions. [19] deals with conditional image synthesis by matching joint distribution. While these works have made some success on joint distribution learning, they either deal with representation learning or emphasize on image synthesis, which can hardly be applied to our noisy label learning problem without special treatment to the noisy labels. Furthermore, they generally lack abilities to learn the correlations among heterogeneous labels of different tasks.

Multi-task learning for facial emotion recognition. Some works utilize auxiliary tasks to facilitate facial emotion recognition [5, 64, 34, 41, 43, 35, 25, 54, 17]. [54] explores

a multi-task learning model for expression recognition and action unit detection by automatically learning the weights for either task. [17] predicts emotions with facial landmark detection as a side task. Recently a large-scale human emotion dataset Aff-Wild2 [25] is proposed with annotations of basic expression classes, valence-arousal values and action unit labels, and multi-task models are developed to facilitate learning for each individual task. However, these works do not treat the labelling noise problem, while noise extensively exists in common facial emotion recognition labels.

3. Problem Formulation

Noisy-labeled facial emotion recognition aims at training a robust model on facial images merely with noisy labels. Let D be the underlying truth distribution generating $(X, Y) \in \chi \times \nu$ pairs from which n iid samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn. After annotating these samples with a certain bias, we obtain corrupted samples $(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$, and let the distribution of $(X, \tilde{Y})$ be D_ρ. The biased annotation model $P(\tilde{Y}|Y)$ is unknown to the learner. Instances are denoted by $x \in \chi \subseteq \mathbb{R}^d$, clean and noisy labels are denoted by $y, \tilde{y}$ respectively.

A traditional way is to model the noise directly with the noisy label distribution definition $p(\tilde{y} = j | y = i) = t_{j,i}$. The probability that an input x is labeled as j in the noisy data can be computed using $t_{j,i}$:

$$p(\tilde{y} = j | x) = \sum_i^c t_{i,j} p(y = i | x), \tag{1}$$

where c is the total class number. In this way, we can modify a classification model using a probability matrix $T = (t_{j,i})$ that modifies its prediction to match the label distribution of the noisy data. Let θ denote parameters of the prediction model, and $\hat{p}(y|x, \theta)$ be the prediction probability of true labels by the classification model. Then the prediction of the final model is given by

$$\hat{p}(\tilde{y} = j | x, \theta, T) = \sum_i^c t_{j,i} \hat{p}(y = i | x, \theta). \tag{2}$$

However, modeling label noise by the transition probability matrix T has several disadvantages: firstly, the transition matrix lacks constraints for the convergence to the true T [50], which can lead optimization to a wrong direction, secondly, the assumption that annotators mislabel one specific class into another certain class in a constant probability may not always hold. Lastly, the probability matrix is a specific case to model discrete labels and fails in the regression task, while continuous noisy labels also exist in many tasks such as affect prediction. Hence, the relationship between the images and the corresponding labels should be considered for a more generalized model which is applicable in different tasks, as well as the multi-task setting.

To address the drawback of the traditional conditional probability modeling, we suggest to remove the transition matrix T and add a constraint on the joint distribution of (X, Y). As noisy labels are generally outliers of the true distribution, our goal is to infer the true joint distribution to reduce impacts of outliers. Distribution-to-distribution supervision as a regularizor makes learning more robust to noise than one-to-one supervision, therefore the key idea is to match two joint distributions. If the aim is achieved, then we are ensured that all marginals as well as all conditional distributions are aligned to some extent. Here, we consider the following two joint probability distributions over two pairs of data and labels:

$$\hat{p}(x, y^0, y^1 = i | \theta) = \hat{p}(x)\hat{p}(y^0, y^1 = i | x, \theta),$$
$$\hat{q}(\tilde{x}, \tilde{y}^0, \tilde{y}^1 = j | \vartheta) = \hat{q}(\tilde{y}^0, \tilde{y}^1 = j)\hat{q}(\tilde{x} | \tilde{y}^0, \tilde{y}^1 = j, \vartheta),$$
$$\tag{3}$$

where (x, y^0, y^1) is a triplet of the input facial image, its predicted non-emotion latent variable, and emotion label, $(\tilde{y}^0, \tilde{y}^1, \tilde{x})$ is a triplet of the random non-emotion latent variable, the noisy emotion label and the facial image inferred from them, and ϑ is the synthesis mapping from $(\tilde{y}^0, \tilde{y}^1)$ to $\tilde{x}$. To achieve a reliable constraint, a natural way is to optimize Kullback–Leibler or Jensen–Shannon divergence for the alignment of the two joint distributions.

In the multi-task setting for facial emotion understanding, labels of different tasks such as categorical emotion labels (discrete emotion classes), and dimensional emotion labels (continuous valence-arousal values) convey complementary information, which can be utilized for better facial emotion recognition. Still, all types of labels are noisy and this joint distribution learning framework can be easily extended to the multi-task setting. Assume each sample X_n is labeled by T types of noisy labels $Y_n^1, \ldots, Y_n^T$, and the goal is to learn the joint distribution of the samples and all corresponding labels $(X, Y^1, \ldots, Y^T)$, then it is desired to optimize the alignment between the following two joint distributions:

$$\hat{p}(x, y^0, y^1, \ldots, y^T | \theta) = \hat{p}(x)\hat{p}(y^0, y^1, \ldots, y^T | x, \theta),$$
$$\hat{q}(\tilde{x}, \tilde{y}^0, \tilde{y}^1, \ldots, \tilde{y}^T | \vartheta) = \hat{q}(\tilde{y}^0, \tilde{y}^1, \ldots, \tilde{y}^T)\hat{q}(\tilde{x} | \tilde{y}^0, \tilde{y}^1, \ldots, \tilde{y}^T, \vartheta),$$
$$\tag{4}$$

where $y^0, \tilde{y}^0$ are non-emotion latent variables.

4. Proposed Method

As it is computationally intractable to model the explicit probability density function of data distributions of real-world data [45], it is generally infeasible to match the two joint distributions with the exact modeling. To overcome this issue, we resort to the generative adversarial modeling methodology [15], which models a distribution with a generator and approximates the model distribution to the true distribution with a discriminator.

To model the two distributions in Eqn. 4, we exploit two components for the generator: One is an encoder G_Y that
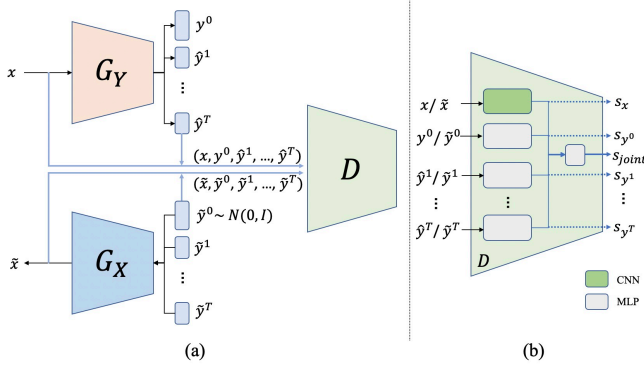
Figure 2. Illustration of (a) the overall architecture, where a pair of symmetric encoder G_Y and decoder G_X aims for simultaneous inference and image generation, with samples from both G_Y and G_X fed into D to align the distributions, (b) the discriminator architecture, which is trained to learn the joint score S_{joint}, as well as the marginal scores S_x, S_{y^i}. x indicates the real input image, $y^0, \hat{y}^1, \ldots, \hat{y}^T$ correspond to the predicted non-emotion and multi-task emotion labels by the encoder G_Y, $\tilde{y}^0$ is a Gaussian noise, $\tilde{y}^1, \ldots, \tilde{y}^T$ are the noisy labels of image x, and $\tilde{x}$ corresponds to the generated image by the decoder G_X.

learns the function θ to infer the clean labels from the input images, while the other is a decoder G_X that learns the function ϑ to produce the facial images with the corresponding expression from the noisy labels. The architecture is illustrated in Fig. 2 (a). As G_X is an image generator that maps from the label space back to the image space, an additional vector y^0 is incorporated into our learning scheme to model other attributes of the image except for the input labels. The encoder G_Y predicts the clean labels $\hat{y}^1, \ldots, \hat{y}^T$ along with the latent noise y^0. In the meantime, the decoder G_X takes a Gaussian noise $\tilde{y}^0$, to generate image $\tilde{x}$ conditioned on the noisy labels $\tilde{y}^1, \ldots, \tilde{y}^T$:

$$G_Y : x \rightarrow (y^0, \hat{y}^1, \ldots, \hat{y}^T),$$
$$G_X : (\tilde{y}^0, \tilde{y}^1, \ldots, \tilde{y}^T) \rightarrow \tilde{x}, \qquad (5)$$

In order to match the joint distributions captured by the encoder and decoder, an adversarial game is played between the generator and the discriminator. In particular, the discriminator (Fig.2 (b)) is designed to match the joint distribution of the group of the facial image, the noise vector, and the multi-task labels from G_Y and G_X. For the joint distribution alignment, a natural way is to feed a pair of the group $(X, Y^0, \ldots, Y^T)$, sampled from the encoder $\hat{p}(x, y^0, \hat{y}^1, \ldots, \hat{y}^T)$ and the decoder $\hat{q}(\tilde{x}, \tilde{y}^0, \tilde{y}^1, \ldots, \tilde{y}^T)$ respectively, into the discriminator network for the adversarial training. Such distribution-to-distribution learning avoids overfitting to specific samples, leading to more robust inference compared to one-to-one supervision which is easily corrupted by the noisy labels.

However, the data within each group is highly heterogeneous and thus the direct concatenation on them could

hurt the distribution learning. To reduce the heterogeneity among the data and multi-task labels, we suggest to exploit multiple network streams to seek their common space where their feature embeddings are less heterogeneous and hence their correlation can be better learned. As shown in Fig. 2 (b), the output embeddings of all the network streams are finally fed into a common network where their joint distributions are learned by the joint score S_{joint}. Besides the advantage to leverage noisy labels of each task, the joint distribution learning also takes advantage of the implicit relationships among related tasks by S_{joint} to facilitate learning for each task. The full objective function is given by

$$\min_{G_X, G_Y} \max_D f(G_Y(x), \tilde{y}) + \lambda(\mathbb{E}_{\hat{p}(x)}[g(D(x, G_Y(x)))] $$
$$+ \mathbb{E}_{\hat{q}(y)}[h(D(G_X(\tilde{y}), \tilde{y}))]), \qquad (6)$$

where $y = (y^0, \hat{y}^1, \ldots, \hat{y}^T)$, $\tilde{y} = (\tilde{y}^0, \tilde{y}^1, \ldots, \tilde{y}^T)$, and the integral forms of the two regularizers $\mathbb{E}_{\hat{p}(x)}[g(D(x, G_Y(x)))]$ and $\mathbb{E}_{\hat{q}(y)}[h(D(G_X(\tilde{y}), \tilde{y}))]$ are given by

$$\mathbb{E}_{\hat{p}(x)}[g(D(x, G_Y(x)))]$$
$$= \int \ldots \int \hat{p}(x)\hat{p}(y|x, \theta)g(D(x, y))dxdy^0 \ldots dy^T, \qquad (7)$$
$$\mathbb{E}_{\hat{q}(y)}[h(D(G_X(\tilde{y}), \tilde{y}))]$$
$$= \int \ldots \int \hat{q}(\tilde{y})\hat{q}(\tilde{x}|\tilde{y}, \vartheta)h(D(\tilde{x}, \tilde{y}))dxd\tilde{y}^0 \ldots d\tilde{y}^T. \qquad (8)$$

The proposed generator and discriminator enable us to optimize the emotion prediction based loss and the distribution match based constraint within a unified framework. On one hand, the encoder G_X is trained to predict the clean labels. On the other hand, the discriminator D learns to align the distributions by distinguishing them, while the encoder G_Y and decoder G_X are trained jointly to fool the discriminator in an adversarial game. According to such scheme, we exploit a min-max objective function for our final solution:

$$\min_{G_X, G_Y} f(G_Y(x), \tilde{y}) + \lambda(\mathbb{E}_{\hat{p}(x)}[\hat{h}(-D(x, G_Y(x)))]$$
$$+ \mathbb{E}_{\hat{q}(y)}[\hat{h}(D(G_X(\tilde{y}), \tilde{y}))])$$
$$\max_D \mathbb{E}_{\hat{p}(x)}[g(D(x, G_Y(x)))] + \mathbb{E}_{\hat{q}(y)}[h(D(G_X(\tilde{y}), \tilde{y}))], \qquad (9)$$

where $y = (y^0, \hat{y}^1, \ldots, \hat{y}^T)$, $\tilde{y} = (\tilde{y}^0, \tilde{y}^1, \ldots, \tilde{y}^T)$, λ plays a trade-off between the multi-task based prediction loss $f(G_Y(x), \tilde{y})$ and the joint distribution learning constraint with the two component functions $g(D(x, G_Y(x)))$, $h(D(G_X(\tilde{y}), \tilde{y}))$ and $\hat{h}(D(G_X(\tilde{y}), \tilde{y}))$, which correspond to adversarial losses [9]. We adopt the hinge loss for the adversarial loss that is commonly used by exiting works like [36, 60]. For the multi-task learning, $f(G_Y(x), \tilde{y})$ can be realized by applying a regular cross entropy loss to the emotion recognition task, with a similarity loss for the continuous labels, which can be replaced by any target-specific loss, such as the Concordance Correlation Coefficient (CCC) loss [44, 52] for affect prediction, which proves

---

Algorithm 1 The proposed method

Require: Batch size m, encoder G_Y, decoder G_X, discriminator D, training iterations n, and hyperparameter λ

1: **for** $i \leftarrow 1$ to n **do**
2:     Sample data $(x_1, \tilde{y}_1^1, \ldots, \tilde{y}_1^T), \ldots, (x_m, \tilde{y}_m^1, \ldots, \tilde{y}_m^T)$ from the dataset
3:     sample Gaussian noise $\tilde{y}_1^0, \ldots, \tilde{y}_m^0$ from $\mathcal{N}(0,1)$
4:     $\tilde{y}_j \leftarrow (\tilde{y}_j^0, \tilde{y}_j^1, \ldots, \tilde{y}_j^T)$ for all j
5:     Update G_X, G_Y:
        $\min_{G_X, G_Y} \frac{1}{m} \sum_{j=1}^m [f(G_Y(x_j), \tilde{y}_j) + \lambda(\hat{h}(-D(x_j, G_Y(x_j))) + \hat{h}(D(G_X(\tilde{y}_j), \tilde{y}_j)))]$
6:     Update D:
        $\max_D \frac{1}{m} \sum_{j=1}^m [g(D(x_j, G_Y(x_j))) + h(D(G_X(\tilde{y}_j), \tilde{y}_j))]$
7: **end for**

---

to be more efficient than the common L2 loss [37, 24, 23]. Formally, the corresponding functions f, g, h, $\hat{h}$ applied for the minimization on the generator are given by

$$f((y_d, y_c), (\tilde{y}_d, \tilde{y}_c)) = L_{CE}(y_d, \tilde{y}_d) + \gamma L_{sim}(y_c, \tilde{y}_c)$$
$$g(z) = \min(0, z - 1),$$
$$h(z) = \min(0, -z - 1),$$
$$\hat{h}(z) = -z,$$

(10)

where $y_d, \tilde{y}_d$ denote the discrete labels, $y_c, \tilde{y}_c$ indicate the continuous labels, and z represents the discriminator's output. L_{CE} and L_{sim} denote the cross entropy loss and the similarity loss for discrete and continuous labels respectively. γ is the trade-off between L_{CE} and L_{sim}.

Although aligning the joint distribution by S_{joint} implicitly matches the marginal distribution, the noise and true label distributions can be very different among noisy heterogeneous labels. In this case, explicit marginal distribution alignment is beneficial. The alignment between the synthetic image distribution and the real image distribution can guide the decoder to generate more realistic images, meanwhile enforcing the distribution of decoder's predicted labels $\hat{y}^1, \ldots, \hat{y}^T$ to match the distribution of the noisy labels $\tilde{y}^1, \ldots, \tilde{y}^T$. Accordingly, each individual network stream within D is expected to learn the corresponding marginal distribution by the marginal scores $\{S_x, S_{y^0}, S_{y^1}, ..., S_{y^T}\}$ as shown in Fig. 2 (b). Since we consider the facial emotion recognition as the target task, we use affect prediction as the auxiliary task to benefit the target task from both the image-to-label relationship and task-to-task relationship. The algorithm of the proposed method is illustrated in Alg. 1.

5. Evaluation

We evaluate the proposed model in two scenarios: (1) a synthetic noisy labeled dataset (CIFAR-10 [26]) for image classification; (2) two practical facial emotion datasets (RAF [28] and AffectNet [37]) for facial emotion recognition. For a more real-world setup, we do not use clean validation labels for model selection, and thus the finally

converged trained model of each comparing method is used directly for evaluation in all experiments. Please refer to supplementary material for more implementation and architecture details.

5.1. Evaluation on Synthetic Noisy Labeled Dataset

Experiment setup. Following [58], the CIFAR-10 dataset [26] for image classification is selected to build the synthetic noisy labeled dataset, as a simulation case to study model behavior with multiple increasing noise. CIFAR-10 includes 60,000 images of size 32x32 in 10 different categories, among which 50,000 are used for training and 10,000 for the test. Images in CIFAR-10 are labeled only for image classes, from which we generate three different sets, in order to simulate our multi-task scenario. Our simulation creates three training sets with different noisy labels, by randomly flipping 20%, 30%, and 40% (each for one set) of the corresponding clean labels. We do not introduce any noise in the test set. The modified noisy labels are uniformly selected across classes. Although this setup is not ideal in the sense of multi-task labels, the proposed model is still applicable, where three inconsistent noisy labels are treated as $\tilde{y}^1, \tilde{y}^2$, and $\tilde{y}^3$, respectively. λ is set as 0.8 in the experiment.

Baselines. As the encoder G_Y is a VGG-backboned network, we compare the proposed model with the following baselines: VGGNet [47] trained on clean labels; VGGNet trained on the majority vote of the three noisy labels; VGGNet trained with all noisy labels; auxiliary image regularizer model (AIR) [1], symmetric cross entropy loss method (SCE) [55]; Co-teaching method [16]; and LTNet [58]. Among the competing methods, LTNet is proposed to deal with inconsistent labels, while other methods mainly tackle the single noisy label issue. We adapted the remaining methods to multiple noisy labels setting, by combining losses of all label sets.

Results and analysis. The results are summarized in Tab. 1. Note that the results of AIR and LTNet are directly from [58], because AIR's result cannot be reproduced by ourselves and LTNet has not released its official code.

Table 1. Test accuracy on CIFAR-10 synthetic dataset.

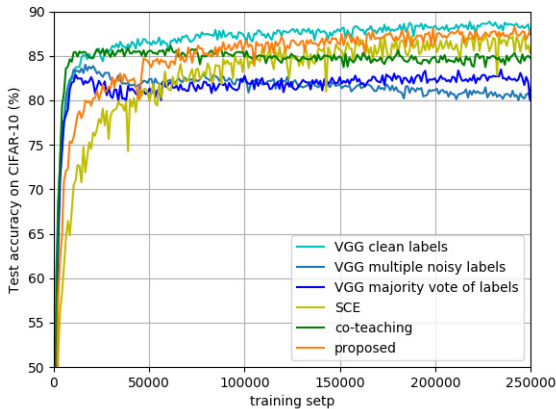| Training data	Model	Test acc (%)
Clean labels	VGGNet [47]	88.55
Multi noisy labels	VGGNet-major vote	82.36
	VGGNet	80.23
	AIR [1]	76.37
	SCE [55]	86.34
	Co-teaching [16]	84.21
	LTNet [58]	87.23
	Proposed	**87.90**

Figure 3. Test accuracy vs. training steps on CIFAR-10 synthetic noisy dataset.

As can be observed, noisy labels severely hurt the learning performance, when used without further treatment. The VGGNet model trained with majority voting label performs poorly because the majority voting only decreases noise ratio, but cannot model the label distribution. The VGGNet model trained with multiple noisy labels also tends to overfit on the noisy training set by one-to-one supervision, leading to a degraded performance in the clean test set. AIR is not trained end-to-end and therefore difficult to optimize. SCE and Co-teaching deal with single noise label issue, from the perspective of robust loss function or complementary networks, hence they lack generalization abilities to the multi-label or multi-task setting. LTNet, which is specifically designed for inconsistent label setting, performs comparable to the VGGNet model trained with clean labels. However, it is not applicable for continuous labels as it models noise by the transition matrix. In comparison, the proposed model trained with multiple noisy labels achieves the best result among compared methods, and is comparable to the model trained using clean labels. Note that our method is not restricted by the number or task types or the noisy labels.

The test accuracy curves of the baselines and the proposed model over training steps are visualized in Fig. 3. Note that the baseline models trained with noisy labels first rise to a peak accuracy quickly and decrease later. In contrast, the test accuracy of the proposed model continues increase over the training steps. Furthermore, we observed that the one-to-one cross entropy loss of the proposed model does not converge to zero, instead settles at a relatively high value. However, the cross entropy loss of VGGNet baseline models converge to almost zero, which indicates overfitting. This observation demonstrates that the joint distribution learning proposed in this paper can avoid the negative influence of the noisy labels.

5.2. Evaluation on Facial Emotion Dataset

The distribution of the synthetic noise in CIFAR-10 dataset is straightforward and only includes one task, while labelling noise can have various patterns in practical emotion datasets due to many reasons (different annotators, labelling protocols, challenges in distinguishing certain expressions, etc.). To study the effectiveness of the proposed model in the emotion recognition task with noisy multi-task labels, we evaluate the model in real-world settings.

Experiment setup. A cross-dataset evaluation protocol is proposed by LTNet [58] to train the model on the combination of AffectNet [37], RAF [28] training set (and some unlabeled facial images), and test on several emotion recognition dataset test sets including AffectNet and RAF. However, such setting is not suitable to our new problem: this training setting assumes that the labelling of AffectNet and RAF are biased and inaccurate, while reporting test accuracy on both AffectNet validation set and RAF test set to demonstrate the superiority of the proposed framework assumes that the same labelling on these two datasets can result in clean labels for test, which is contradictory with the training. Therefore, for the suggested new problem, we propose a more appropriate evaluation procedure on RAF and AffectNet. We use machine labeling by pretrained models, which inherently introduces noise due to domain gap and human-machine disagreements.

Baselines. As existing methods for noisy label learning either are not applicable, or cannot be easily adapted to the multi-task setting or continuous labels, we choose to train two of state-of-the-art methods for single discrete noisy label learning, i.e., SCE [55], Co-teaching [16], only with discrete expression labels. For comparison, we also train the VGGNet model and a degraded version of our model in the single-task setup only with discrete expression labels or continuous valence-arousal labels. To evaluate the proposed multi-task model's full effectiveness, we train it with both expression and valence-arousal labels. With no existing work applicable for multi-task noisy label learning with both discrete and continuous labels, we train another VGGNet model with multi-task labels with loss combination of both tasks for comparison.

The **RAF** dataset consists of 15,339 real-world facial

Table 2. Evaluation results on facial emotion datasets. Single-task refers to models trained only with categorical expression labels or valence-arousal labels, and multi-task refers to models trained with both expression and valence-arousal labels. Emotion/Acc (%) denotes test accuracy of categorical expression prediction, the higher the better. VA/CCC, VA/MSE denote CCC and MSE metrics of valence-arousal prediction respectively, the higher the better for CCC and the lower the better for MSE. (**Bold**: best, <u>Underline</u>: second best)

Setting	Model	RAF-base	AffectNet-base		
Task/Metric		Expression/Acc (%)	Expression/Acc (%)	VA/CCC	VA/MSE
Single-task	VGGNet [47]	72.64	43.42	0.6254	0.1438
	SCE [55]	73.96	42.87	-	-
	Co-teaching [16]	<u>75.43</u>	42.36	-	-
	Proposed	74.02	<u>44.52</u>	<u>0.6354</u>	<u>0.1284</u>
Multi-task	VGGNet	73.15	43.36	0.6263	0.1354
	Proposed	**76.10**	**46.08**	**0.6727**	**0.1248**

images collected from Flickr, including 12,271 images for training and 3,068 images for testing. Each image was labelled as one of the seven basic facial emotions (i.e., neutral, happy, sad, surprise, angry, fear, disgust) by about 40 independent annotators, followed by an EM algorithm to assess the reliability of each annotator. The **AffectNet** dataset is a multi-task in-the-wild dataset including around 450,000 training images, and 5,500 validation images. Each image in AffectNet is labeled by two labels: one of the eight discrete expression classes (seven basic expressions with an additional contempt class), and the continuous valence-arousal values. The labels for the test set is not public therefore we only report accuracy results on the validation set, which is not seen during training. We select images with the seven basic expression labels in the experiment, resulting in 283,910 training images and 3,500 test images.

We create two scenarios for evaluation: (1) the RAF training set is relabeled for both expression classes and valence-arousal values by an AffectNet pretrained model for training, and we keep RAF test set unchanged for evaluation; (2) the expression classes of AffectNet training set is relabeled by a RAF pretrained model for training, together with the original valence-arousal labels of AffectNet, and the trained model is evaluated on the AffectNet validation set. We denote the two experiment scenarios by **RAF-base** and **AffectNet-base** respectively. For valence-arousal prediction, we only report results in AffectNet-base case since no human valence-arousal labels are available on RAF.

Results and analysis. The experiment results on facial emotion datasets are illustrated in Tab. 2. For the discrete expression prediction, in the single-task RAF-base case, SCE and the proposed model achieve similar performance, all higher than VGGNet, and Co-teaching achieves the best performance. Co-teaching trains two networks simultaneously, with each one selecting small-loss instances (more likely to have correct labels) and teaching its peer network. However, such instances are less likely to occur with in-

creasing noise intensity. Therefore, Co-teaching is a strong baseline for the single-task RAF-base case, where the noise level is relatively low. For the AffectNet-base case where the model is exposed to higher noise, SCE and Co-teaching face difficulties for optimization with more complicated noise distributions, and perform even worse than VGGNet. Moreover, it is non-trivial to adapt these two baselines to multi-task. In contrast, the proposed model takes full advantage from multi-task noisy labels and achieves the best performance. For the continuous valence-arousal prediction, we utilize the commonly used Concordance Correlation Coefficient (CCC) and Mean Square Error (MSE) as the evaluation metrics. We can observe that the proposed model trained in the multi-task setting significantly improves prediction performance compared with the VGGNet baseline method and the single-task models.

For the case of clean labels, it is natural to combine loss functions of each task in multi-task learning [25]. Nevertheless such simple loss combination is not necessarily beneficial in the presence of noisy labels for each task, or even harmful, which could be observed in Tab. 2. However, the proposed multi-task model demonstrates superior performance compared with multi-task VGGNet and single-task models. Our joint distribution learning scheme not only leverages noise for each task individually, but also learns a robust correlation between different tasks as well as the image, to take advantage of labels of both tasks and utilize richer information for emotion recognition without being corrupted by the noise in labels of each task.

5.3. Ablation Study

Here we conduct the ablation study to investigate the contributions of marginal scores $\{S_x, S_{y^0}, S_{y^1}, ...\}$ and joint scores S_{joint} in D, the decoder G_X and extra labels of the proposed model. The ablation study is implemented on the following four cases: (1) single-label model on the synthetic noisy CIFAR-10 dataset with 20% noise in the

Table 3. Ablation study for different components of the proposed model: w/o joint, w/o marginal and w/o G_X respectively remove S_{joint}, $\{S_x, S_{y^0}, S_{y^1}, ...\}$ and G_X in Fig. 2.

Setting	VGGNet	w/o joint	w/o marginal	w/o G_X	Proposed
Single-label CIFAR-10*	77.82	77.93	83.65	83.51	**84.78**
Multi-label CIFAR-10	80.23	79.10	85.48	87.88	**87.90**
Single-task RAF-base	72.64	72.28	73.19	73.37	**74.02**
Multi-task RAF-base	73.15	73.24	74.40	74.46	**76.10**

* Use one single noisy label set with 20% noise for training

(a)   (b)   (c)

Figure 4. Samples synthesized by the decoder G_X conditioned on the input label: (a) samples from CIFAR-10, and each column conditioned on input class labels as *airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck* from left to right; (b)(c) samples from RAF/AffectNet, respectively, and each column conditioned on input expression labels as *neutral, happy, sad, surprise, fear, disgust, anger* from left to right. The rotations and boundaries are due to data augmentation during training.

training labels, (2) multi-label model on the synthetic noisy CIFAR-10 dataset, (3) the single-task model on RAF-base case, and (4) the multi-task model on RAF-base case). Results are shown in Tab. 3. With respect to design of D, only marginal distribution matching (by $\{S_x, S_{y^0}, S_{y^1}, ...\}$) without joint distribution learning has marginal benefit (w/o joint), while the joint distribution learning by S_{joint} in D is essential to combat the noise labels (w/o marginal). Additional marginal scores on top of S_{joint} brings extra improvements by explicit marginal distribution matching (proposed). The decoder G_X is beneficial by facilitating the joint distribution learning in a more balanced way. Here the non-emotion variable y^0 in Fig. 2 is necessary to encode other image attributes so that G_X has sufficient information for generation. Training with multiple labels further enhances the model by capturing the correlation among complementary labels, which is validated both in the CIFAR-10 case and practical emotion learning scenario.

5.4. Conditional Image Synthesis

Our proposed joint distribution learning framework learns inference and conditional image synthesis simultaneously. Since the distribution learning only serves as a regularizar in our algorithm, the generated images are not necessarily of the highest quality, but should carry explicit semantic meaning. Fig. 4 presents generated samples on CIFAR-10 (Fig. 4 (a)), RAF (Fig. 4 (b)) and AffectNet (Fig. 4 (c)). The decoder G_X can generate correct images given the conditional label, which demonstrates that the two joint distributions could be aligned with the adversarial training.

6. Conclusion and future work

This paper introduces an interesting problem of facial emotion recognition with noisy multi-task annotations, which has a high potential to reduce human labelling efforts for multi-task learning. To better treat the problem, we introduce a new formulation from the view of joint distribution match. Following the suggested formulation, we exploit a new adversarial learning method to jointly optimize the emotion prediction and the joint distribution learning. Finally we study the setup of synthetic noisy labeled dataset and practical noisy multi-task datasets, and experiments demonstrate the clear advantage of the proposed method for the new problem. While we can roughly setup the trade-off between the emotion prediction loss and the joint distribution match based constraint, automatically adapting the balance would be interesting to study in our future work.

References

[1] Samaneh Azadi, Jiashi Feng, Stefanie Jegelka, and Trevor Darrell. Auxiliary image regularization for deep cnns with noisy labels. *arXiv preprint arXiv:1511.07069*, 2015.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[3] Xi Chen, Qihang Lin, and Dengyong Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *International conference on machine learning*, pages 64–72, 2013.

[4] LI Chongxuan, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *Advances in neural information processing systems*, pages 4088–4098, 2017.

[5] Terrance Devries, Kumar Biswaranjan, and Graham W Taylor. Multi-task learning of facial landmarks and expression. In *2014 Canadian Conference on Computer and Robot Vision*, pages 98–103. IEEE, 2014.

[6] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112. IEEE, 2011.

[7] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[8] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pages 10541–10551, 2019.

[9] Hao-Wen Dong and Yi-Hsuan Yang. Towards a deeper understanding of adversarial losses. *arXiv preprint arXiv:1901.08753*, 2019.

[10] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[11] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[12] Zhe Gan, Liqun Chen, Weiyao Wang, Yuchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. Triangle generative adversarial networks. In *Advances in neural information processing systems*, pages 5247–5256, 2017.

[13] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

[14] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[16] 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 *Advances in neural information processing systems*, pages 8527–8537, 2018.

[17] Guosheng Hu, Li Liu, Yang Yuan, Zehao Yu, Yang Hua, Zhihong Zhang, Fumin Shen, Ling Shao, Timothy Hospedales, Neil Robertson, et al. Deep multi-task learning to recognise subtle facial expressions of mental states. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018.

[18] Ninghang Hu, Gwenn Englebienne, Zhongyu Lou, and Ben Kröse. Learning to recognize human activities using soft labels. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):1973–1984, 2016.

[19] Ayush Jaiswal, Wael AbdAlmageed, Yue Wu, and Premkumar Natarajan. Bidirectional conditional generative adversarial networks. In *Asian Conference on Computer Vision*, pages 216–232. Springer, 2018.

[20] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.

[21] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6-7):907–929, 2019.

[24] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[25] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019.

[28] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[29] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[30] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1805–1812, 2014.

[31] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[32] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. *arXiv preprint arXiv:1806.02612*, 2018.

[33] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". In *Advances in Neural Information Processing Systems*, pages 960–970, 2017.

[34] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE, 2017.

[35] Zuheng Ming, Junshi Xia, Muhammad Muzzamil Luqman, Jean-Christophe Burie, and Kaixing Zhao. Dynamic multi-task learning for face recognition with facial expression. *arXiv preprint arXiv:1911.03281*, 2019.

[36] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[37] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[38] Maja Pantic. Facial expression recognition. In *Encyclopedia of biometrics*, pages 400–406. Springer, 2009.

[39] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[40] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[41] Gerard Pons and David Masip. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*, 2018.

[42] Yunchen Pu, Shuyang Dai, Zhe Gan, Weiyao Wang, Guoyin Wang, Yizhe Zhang, Ricardo Henao, and Lawrence Carin. Jointgan: Multi-domain joint distribution learning with generative adversarial nets. *arXiv preprint arXiv:1806.02978*, 2018.

[43] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017.

[44] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1335–1336, 2015.

[45] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455, 2009.

[46] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

[49] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.

[50] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019.

[51] Bruce Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[52] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.

[53] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France, 2010.

[54] Chu Wang, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Multi-task learning of emotion recognition and facial action unit detection with adaptively weights sharing network. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 56–60. IEEE, 2019.

[55] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 322–330, 2019.

[56] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7017–7025, 2019.

[57] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement

help generalization against label corruption? *arXiv preprint arXiv:1901.04215*, 2019.

[58] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018.

[59] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008.

[60] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[61] Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2019.

[62] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pages 8778–8788, 2018.

[63] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[64] Rui Zhao, Tianshan Liu, Jun Xiao, Daniel PK Lun, and Kin-Man Lam. Deep multi-task learning for facial expression recognition and synthesis based on selective feature sharing. *arXiv preprint arXiv:2007.04514*, 2020.

[65] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016.

[66] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.