Facial Expression Recognition with Inconsistently Annotated Datasets

Jiabei Zeng[1], Shiguang Shan[1,2,3], and Xilin Chen[1,2]

Key Lab of Intelligent Information Processing of Chinese Academy of Sciences
(CAS), Institute of Computing Technology, CAS, Beijing 100190, China[1]
University of Chinese Academy of Sciences, Beijing 100190, China[2]
CAS Center for Excellence in Brain Science and Intelligence Technology[3]
{jiabei.zeng, sgshan, xlchen}@ict.ac.cn

Abstract. Annotation errors and bias are inevitable among different facial expression datasets due to the subjectiveness of annotating facial expressions. Ascribe to the inconsistent annotations, performance of existing facial expression recognition (FER) methods cannot keep improving when the training set is enlarged by merging multiple datasets. To address the inconsistency, we propose an Inconsistent Pseudo Annotations to Latent Truth(IPA2LT) framework to train a FER model from multiple inconsistently labeled datasets and large scale unlabeled data. In IPA2LT, we assign each sample more than one labels with human annotations or model predictions. Then, we propose an end-to-end LTNet with a scheme of discovering the latent truth from the inconsistent pseudo labels and the input face images. To our knowledge, IPA2LT serves as the first work to solve the training problem with inconsistently labeled FER datasets. Experiments on synthetic data validate the effectiveness of the proposed method in learning from inconsistent labels. We also conduct extensive experiments in FER and show that our method outperforms other state-of-the-art and optional methods under a rigorous evaluation protocol involving 7 FER datasets.

1 Introduction

Facial expressions convey varied and nuanced meanings. Automatically recognizing facial expression is important to understand human's behaviors and interact with them. During the last decades, the community has made promising progresses in building datasets and developing methods for facial expression recognition (FER). Datasets have sprung up for both in-the-lab and in-the-wild facial expressions, such as CK+[20], MMI[28], Oulu-CASIA[33], SFEW/AFEW[7], AffectNet[22], EmotioNet[2], RAF-DB[16], and others. Based on these datasets, lots of FER approaches are proposed and achieve the state-of-the-art performance[27, 25, 30, 4, 14, 34, 18].

However, errors and bias of human annotations exist among different datasets. As been known, it is subjective to classify the face expression into several emotional categories. Human's understanding of facial expressions varies with different cultures, living environments, and other experiences. Although the human

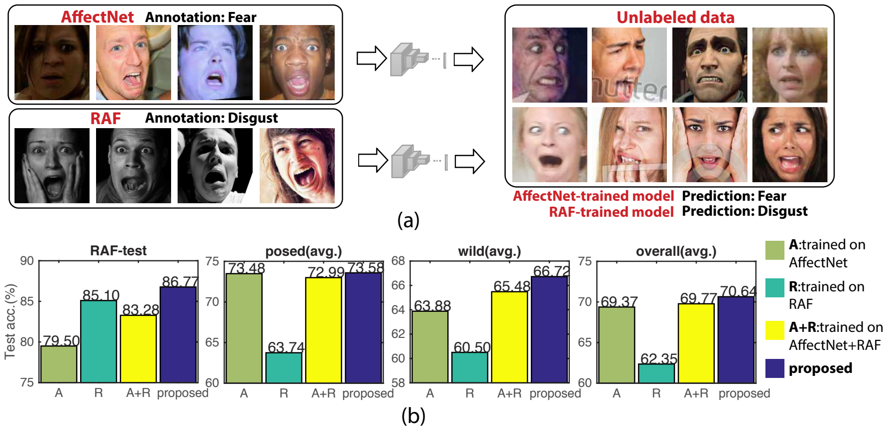


Fig. 1. (a) Inconsistent predictions due to the annotations bias in AffectNet and RAF. (b) Test accuracy on different datasets with varied combination of training data.

coders are claimed to be trained before the annotation tasks[16, 30], the bias of annotations is inevitable among different datasets, because teams from different backgrounds would have different criterions in controlling the quality of the released datasets. The annotation bias results in two main issues: (1) FER systems are easy to heritage the recognition bias from the training dataset. Fig.1(a) shows examples of the inconsistent predictions caused by the annotation bias in AffectNet and RAF datasets. The faces presented on the left have similar expressions but they are labeled as "fear" in AffectNet and as "disgust" in RAF. As a consequence, the two models trained from AffectNet and RAF have inconsistent predictions on the unlabeled images presented on the right. They are predicted as "fear" by the AffectNet-trained model but as "disgust" by the RAF-trained one. (2) It is difficult to accumulate the benefit of different datasets by simply merging them as a whole during the training process. Fig.1(b) shows the test accuracy on different test sets with varied combination of training data. As can be seen, models trained from the most data are not sufficient to be the best one. On RAF-test, the model trained from the union of AffectNet and RAF(A+R) has lower test accuracy than the one trained from RAF only. On posed facial expression data, model A+R performs worse than the one from AffectNet only.

To address the issues, we propose a 3-step framework to build a FER system on inconsistently annotated datasets. We name the framework as Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) because it tags multiple labels for each image with the human annotations or predicted pseudo labels, and then learns a FER model to fit the latent truth from the inconsistent pseudo labels. Fig.2 illustrates the main idea of the IPA2LT framework. IPA2LT consists of three steps. It first trains two machine annotators from data A and B respectively. The predictions by machine annotators and the human annotations are probably to be inconsistent. They are used as multiple labels for each image in

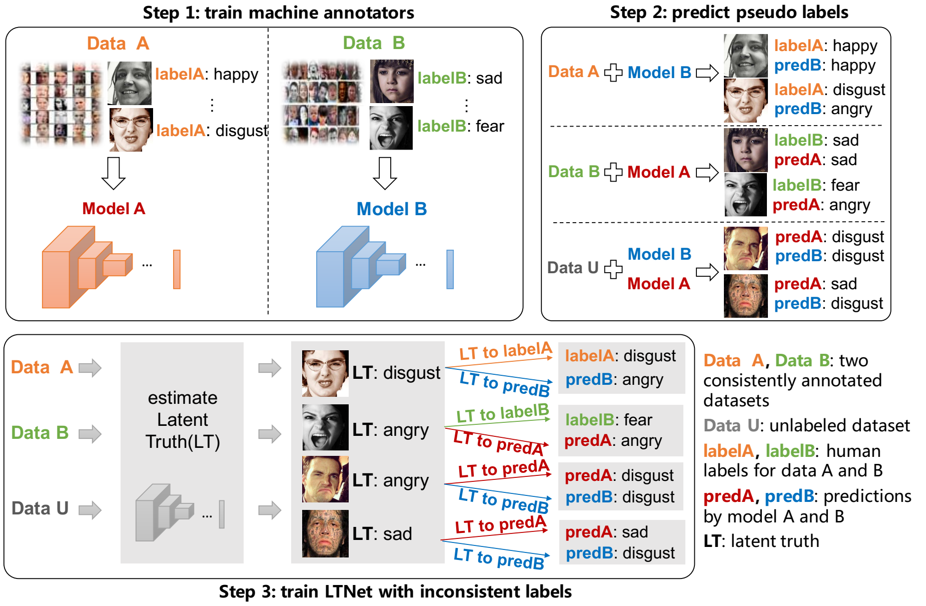


Fig. 2. Three steps in the proposed Inconsistent Pseudo Annotations to Latent Truth(IPA2LT) framework.

several human labeled datasets as well as the large scale unlabeled data. Unlabeled data serves as the bridge between data A and B by sharing the same machine annotators with them. Then, IPA2LT trains a Latent Truth Net(LTNet) to discover the latent true label. LTNet is end-to-end trainable and therefore it can estimate the latent truth depending on both the input face image and the inconsistent labels. During the inference, the learned LTNet is applied to estimate the true label for a new face. Our contributions are summarized as follows:

1. We propose a relatively unexplored problem: how to learn a classifier from more than one datasets with different annotation preferences. To the best of our knowledge, it is the first work that addresses the annotation inconsistency in different FER datasets.

2. We introduce a IPA2LT framework to train a FER model from multiple inconsistently labeled datasets and the large scale unlabeled data. In the framework, we propose an end-to-end trainable LTNet[1] embedded with a scheme of discovering the latent truth given multiple observed(or predicted) labels and the input face images.

3. Experiments on both synthetic and real data validate the effectiveness of the proposed method in learning from inconsistent labels. We conduct extensive experiments in FER and show the advantages of IPA2LT over the state-of-the-art under a rigorous evaluation protocol involving 7 FER datasets.

[1] Code available at https://github.com/dualplus/LTNet.

2 Related work

The proposed method aims to train a classifier from inconsistently labeled datasets. The inconsistent labels associate with multiple noisy labels. Therefore we review the related work about the methods with inconsistent labels and noisy labels.

Methods with inconsistent labels: A straightforward way to address inconsistent labels is using "soft labels" during the training process. For example, He et. al.[11] dealt with the noisy labels from multiple annotators by proposing a loss that incorporates the "soft labeling" in a max-margin learning framework. The soft labels based methods assume all the annotations to have equal contributions in rating the importance, and ignore that some annotations are less reliable than others.

Another typical way is the ones in estimating the ground truth in crowdsourcing[35]. These works estimate the latent truth from different annotators using EM algorithm. As early as 1979, Dawid and Skene[5] proposed to solve the labeling task with k different categories by assuming each worker to be associated with a $k \times k$ confusion matrix, where the (l, c)-th entry represents the probability that a sample in class l is labeled as class c by the worker. The EM-based methods have had empirical success in determining the labels in crowdsourcing[3, 19, 36, 32]. Considering the label qualities from different annotators, methods are proposed to iteratively qualify the annotators and estimate the latent truths, such as using Gaussian Mixture Model and Bayesian Information Criterion[31], Chinese restaurant process[23], and other probabilistic frameworks.

However, the methods in crowdsourcing focus on estimating the ground truth of the samples that already have a set of inconsistent annotations. They ignore the mapping between the latent truth and the input data and make few efforts on learning a predictor to estimate labels for unseen samples. We focus on training the predictor for unseen samples and capture the relations between the input data and the true labels.

Methods with noisy labels: To address the noisy labels, numbers of methods were proposed. One idea is to leverage a small set of clean data. The clean data is used to assess the quality of the labels during the training process [17, 29, 6], or to train the feature extractors[1], or to estimate the distribution of noisy labels[26]. For example, Li et.al.[17] proposed a unified distillation framework using information from a small clean dataset and label relations in knowledge graph, to hedge the risk of learning from noisy labels. Veit et.al.[29] comprised a multi-task network that jointly learns to clean the noisy annotations and to classify the images. Azadi et.al.[1] selected reliable images by an auxiliary image regularization for deep CNNs with noisy labels. The CNN feature extractor was trained from a set of clean data. Sukhbaatar and Fergus [26] introduced an extra layer into the network to adapt the network outputs to match the noisy label distribution and they estimated the layer's parameters from clean and noisy data.

Other methods do not need a set of clean data but assume extra constrains or distributions on the noisy labels[21], such as proposing losses for randomly flipped labels[24], regularizing the deep networks on corrupted labels by a Metor-

Net[13], augmenting the prediction objective with the similarities and improving the learner iteratively using bootstrapping[15], and other approaches that introducing constrains. As a very similar work to the proposed LTNet, Goldberger and Ben-Reuven[9] modeled the noise by a softmax layer that connects the correct labels to the noisy ones. They presented a neural-network approach that optimizes the same likelihood function as optimized by the EM algorithm. LTNet differs from this work, as well as other methods with noisy labels, by that we consider each sample having several annotations rather than one for each. Therefore, we can discover the noise patterns from the multiple anotations.

3 Proposed method

3.1 IPA2LT framework

We propose an Inconsistent Pseudo Annotations to Latent Truth(IPA2LT) framework to train a FER model from multiple inconsistently labeled datasets. IPA2LT leverages large scale unlabeled data as well as several human labeled datasets. In IPA2LT, each sample has more than one annotations, including the observed or predicted ones. With the inconsistent pseudo annotations, IPA2LT builds an end-to-end network LTNet to fit the latent truth.

Figure 2 illustrates the 3-step IPA2LT framework. Let us suppose that we are given two human labeled datasets A and B, and the unlabeled data U. Note that the IPA2LT framework is flexible to be adapted to more than two human labeled datasets. As can be seen in the Step 1 in Fig. 2, IPA2LT trains two machine coders ($\mathrm{M_A}$ and $\mathrm{M_B}$) from the two datasets A and B, respectively. In Step 2, IPA2LT makes pseudo annotations for both the human labeled and unlabeled data using the predictions by machine coders. Specifically, we predict data A using $\mathrm{M_B}$ and thus data A has two sets of labels, i.e., the human annotated one and the $\mathrm{M_B}$-predicted one. Similarly, data B has two sets labels as the human annotated one and the $\mathrm{M_A}$-predicted one. We also estimate two sets of labels for the large scale unlabeled data U using $\mathrm{M_A}$ and $\mathrm{M_B}$, respectively. Then, each sample has two labels that are probably inconsistent. In Step 3, IPA2LT trains an end-to-end Latent Truth Net(LTNet) to discover the latent truth considering the inconsistent labels and the input images. A scheme of discovering the latent truth is embedded in LTNet. During the inference, the learned LTNet can be used to estimate the true label for a new face image.

The first two steps can be complemented easily by adopting any classification methods as the machine coders and using them to predict the pseudo labels. Yet, it is non-trivial to train a model that fits the latent truth provided multiple inconsistent annotations. To achieve this, we propose an end-to-end trainable LTNet that is embedded with a scheme of discovering the latent truths from multiple observed(or predicted) labels and the input images.

3.2 Formulation of LTNet

Inconsistent annotations are caused by the labeling preference bias of different annotators when they are labeling a set of data. Each annotator has a coder-specific bias in assigning the samples to some categories. Mathematically speaking, let $\mathcal{X} = \{\mathbf{x}_i, \ldots, \mathbf{x}_N\}$ denote the data, $\mathbf{y}^c = [y_1^c, \ldots, y_N^c]$ the annotations by coder c. Inconsistent annotations assume that

$$P(y_n^i|\mathbf{x}_n) \neq P(y_n^j|\mathbf{x}_n), \forall \mathbf{x}_n \in \mathcal{X}, i \neq j \tag{1}$$

where $P(y^i|\mathbf{x}_n)$ denotes the probability distribution that coder i annotates sample $\mathbf{x}_n$.

LTNet assumes that each sample $\mathbf{x}_n$ has a latent truth y_n. Without the loss of generality, let us suppose that LTNet classifies $\mathbf{x}_n$ into category i with probability $P(y_n = i|\mathbf{x}_n; \varTheta)$, where $\varTheta$ denotes the network parameters. If $\mathbf{x}_n$ has a ground truth of i, coder c has an opportunity of $\tau_{ij}^c = P(y_n^c = j|y_n = i)$ to annotate $\mathbf{x}_n$ as j, where y_n^c is the annotation of sample $\mathbf{x}_n$ by coder c. Then, the sample $\mathbf{x}_n$ is annotated as label j by coder c with a probability of:

$$P(y_n^c = j|\mathbf{x}_n; \varTheta) = \sum_{i=1}^{L} P(y_n^c = j|y_n = i) P(y_n = i|\mathbf{x}_n; \varTheta), \tag{2}$$

where L is the number of categories and $\sum_j^L P(y_n^c = j|y_n = i) = \sum_j^L \tau_{ij}^c = 1$.

Given the annotations from C different coders on data $\mathcal{X}$, LTNet aims to maximize the loglikelihood of the observed annotations as:

$$\max_{\varTheta, \mathbf{T}^1, \cdots, \mathbf{T}^C} \log\left(P(\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^C|\mathcal{X}; \varTheta)\right), \tag{3}$$

where $\mathbf{y}^c = [y_1^c, y_2^c, \cdots, y_N^c]^\top$ is the annotations by coder c on the N samples in $\mathcal{X}$. $\mathbf{T}^c = [\tau_{ij}^c]_{L\times L}$ denotes the transition matrix with rows summed to 1. The loglikelihood is computed as:

$$\begin{aligned} \log\left(P(\mathbf{y}^1, \cdots, \mathbf{y}^C|\mathcal{X}; \varTheta)\right) &= \log\left(\prod_{n=1}^{N}\prod_{c=1}^{C} P(y_n^c|\mathbf{x}_n; \varTheta)\right) \\ &= \sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{j=1}^{L} \mathbf{1}(y_n^c = j) \log\left(\tau_{ij}^c P(y_n = i|\mathbf{x}_n; \varTheta)\right) \end{aligned} \tag{4}$$

where $\mathbf{1}(\cdot)$ is the indicating function. It equals to 1 if the condition in the bracket holds and equals to 0 otherwise.

3.3 Solutions to the objective function of LTNet

The objective function (3) aims to find the transition matrics $\mathbf{T}^1, \cdots, \mathbf{T}^c$ and the optimal parameters $\varTheta$ that are used to compute the latent truths for the input data $\mathcal{X}$.

It is difficult to optimize (3) because it is NP hard. An intuitive approach is to solve (3) in two separate steps: estimate the latent truth using Dawid&Skene's EM algorithm[5] and then train the network with the estimated labels. The EM

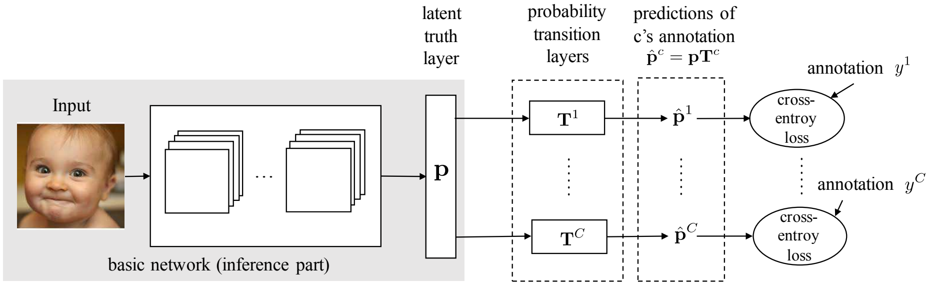


Fig. 3. Architecture of the end-to-end trainable LTNet. Each row of the transition matrix $\mathbf{T}$ is constrained to be summed to 1.

algorithm alternatively optimizes the latent truth $\{y_n\}_{n=1}^N$ and the transition matrices $\mathbf{T}^c, \forall c \in \{1, \cdots, C\}$ by maximizing:

$$\max_{\{y_n\}_{n=1}^N, \mathbf{T}^1, \cdots, \mathbf{T}^C} \log P(\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^C), \tag{5}$$

where $P(\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^C) = \prod_{n=1}^N \prod_{c=1}^C \prod_{j=1}^C \left(\tau_{ij}^c P(y_n = i)\right)^{\mathbf{1}(y_n^c = j)}$. During the E-step in each iteration of EM algorithm, we fix the transition matrices $\mathbf{T}^c, \forall c \in \{1, \cdots, C\}$ and compute the expectation of latent truth $\{y_n\}_{n=1}^N$. During the M-step, we fix the latent truth $\{y_n\}_{n=1}^N$ and optimize the transition matrices $\mathbf{T}^c, \forall c \in \{1, \cdots, C\}$. After several iterations in EM algorithm, we can have the estimated latent truth for each sample. Then, we train a convolution neural network for FER, whose parameters are Θ, to fit the estimated latent truth.

The 2-step solution estimates the latent truth and learns the classifier parameters separately. It ignores the relations between the input images and the latent truths. The latent truth should also be determined according to the raw images rather than only to the annotations by multiple coders. To this end, we integrate the Dawid&Skene's[5] and the CNN into an end-to-end trainable architecture LTNet.

Figure 3 illustrates the architecture of LTNet. LTNet takes facial images as inputs and estimates the latent truths' probability distribution $\mathbf{p}$ through a basic deep convolution neural network. Then, rather than minimizing the discrepancy between the estimated truths and the observed labels directly, LTNet predicts each coder's annotation and minimizes the discrepancy between the predicted and observed annotations. Specifically, the estimated truths are passed through a coder-specific probability transition layer to get the predictions of coder c's annotation. Coder c's probability transition layer has the transition matrix $\mathbf{T}^c \in \mathbb{R}^{L \times L}$ as parameters, where L is the number of categories. $\mathbf{T}^c$'s entry τ_{ij}^c denotes the probability that coder c annotates a sample as category j if the sample is with ground truth i. Each row of $\mathbf{T}^c$ indicates a probability distribution and thus is summed to 1. The probability transition layer takes the input as the ground truths' probability $\mathbf{p} = [P(y = 1|\mathbf{x}, \Theta), \cdots, P(y = L|\mathbf{x}, \Theta)]^\top$, and then outputs the predicted distribution of coder c's annotation as $\hat{\mathbf{p}}^c = \mathbf{p}^\top \mathbf{T}^c$. To ensure that each row of $\mathbf{T}^c$ is summed to 1, we normalize each row of $\mathbf{T}^c$ before each forward process. Note that other tricks can be adopted to keep $\mathbf{T}^c$'s rows summed to 1 as well. For example, the probability transition layers can take the row of $\mathbf{T}^c$ as the output of soft-max operation on a L-dimensional vector.

Finally, parameters in LTNet is learned by minimizing the cross-entropy loss of the predicted and observed annotations for each coder as:

$$\min_{\Theta,\{\mathbf{T}^1,\cdots,\mathbf{T}^C\}} \quad -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{L}\mathbf{1}(y_n^c = k)\log(\hat{p}_n^c(k)) \tag{6}$$

$$s.t. \quad \sum_{j}^{L}\tau_{ij}^c = 1, \forall i = 1,\ldots,L \tag{7}$$

where N is the number of samples, C is the number of coders, and L is the number of categories. τ_{ij}^c is the element of $\mathbf{T}^c$. y_n^c is the annotation of the n-th sample by coder c. $\hat{\mathbf{p}}_n^c = [\hat{p}_n^c(1),\cdots,\hat{p}_n^c(L)]^\top$ denotes the predicted distribution of coder c's annotation on the n-th sample. Solving (6) is equivalent to solving the objective function (3). LTNet can be optimized by back-propagation methods.

4 Experiments

4.1 Evaluations on synthetic inconsistently labeled data

Data The synthetic data was builded from the widely used CIFAR-10 dataset, which contained 60,000 tiny images in 10 categories. In CIFAR-10, 10000 images (1000 image/category) were chosen as the test part and the others were the training part. We synthesized 3 pieces of inconsistent annotations for the training samples by randomly revising 20%, 30%, and 40% of the corrected labels, respectively. The artificial noisy labels were distributed uniformly in different categories. The test set remained clean and was used to evaluate the approaches in our experiments.

Comparison to other methods We compared LTNet with 3 types of methods: i) basic CNNs trained on a single set of noisy labels; ii) basic CNNs trained on all the 3 pieces of noisy labels with different label selecting strategy, i.e., simply mixing all the labels or selecting the majority ratings as labels; and iii) state-of-the-art methods that address inconsistent or noisy labels, i.e., AIR[1], NAL[9], EM+CNN[5, 32]. In AIR, we trained a CNN from the mixture of the noisy labels and used the features from the trained CNN to do the afterward L_{12}-norm regularization. In NAL, we regarded the mixture of the three noisy sets as a whole. EM+CNN is similar to the 2-step solution in Section 3.3, where we used EM algorithms to estimate the latent truth, and then trained a CNN on the latent truth. In our experiments, we used two ways to initialize the EM algorithm, i.e., majority rating [5] and spectral method[32]. The source code for AIR and EM are downloaded from the authors' website. NAL was re-implemented by ourselves. No other datasets were used to pre-train or initialize the models in all of the experiments.

Table 1. Test accuracy on CIFAR-10 with noisy labels

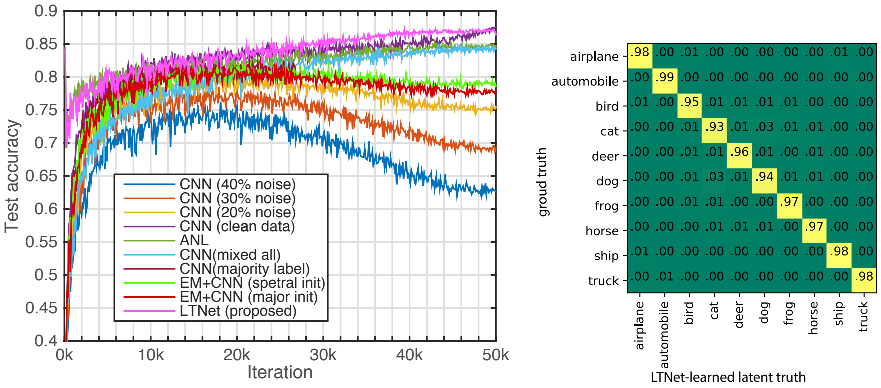
Training data	Methods	Test Acc.(%)
clean	*basic CNN*	*87.43*
40% noise (A)	basic CNN	62.50
30% noise (B)	basic CNN	68.70
20% noise (C)	basic CNN	75.15
mixture of ABC	basic CNN	84.04
mixture of ABC	AIR[1]	76.37
mixture of ABC	NAL[9](re-implementation)	84.41
majority ratings of ABC	basic CNN	82.95
inconsistent annotations	EM+CNN(majority init.[5])	77.65
inconsistent annotations	EM+CNN(spectral init.[32])	78.82
inconsistent annotations	LTNet (proposed)	**87.23**

The test accuracy of all the methods are shown in Table 1. We also report the test accuracy of the basic CNN trained on clean data. As can be seen in Table 1, whichever methods are used, using all the inconsistently labeled sets boosts the performance of the models using a single noisy set. Because the multiple annotations, although being inconsistent, convey more correct information than a single set of noisy labels.

Within the methods trained on mixture data, we observe that the end-to-end methods (e.g., basic CNN on mixture data or majority ratings, NAL, LTNet) are significantly better than the step-by-step methods (e.g., AIR, EM+CNN). A viable explanation is that the end-to-end methods can intrinsically capture the relations between the input image and inconsistent labels. But the step-by-step methods separately capture the relations between the input images and the estimated labels, and the relations between the latent truths and inconsistent labels. Among all the end-to-end methods, the proposed LTNet achieves the highest test accuracy and has a comparable performance to the CNN trained from clean data.

To further investigate the methods, we plot the test accuracy curve during the training iterations in Fig. 4(a). The x-axis is the iteration number during the training process. As can be seen, the test accuracy curves of LTNet, CNN(clean data), CNN(mixed all), and ANL keep increasing during the training, while those of CNN (with 40%, 30%, or 20% noisy labels) and EM+CNN(spectral or major init) reach a peak value and then decrease as the training iterates. Because the latter methods are unable to distinguish the incorrect label information from the noisy labels or estimated ground truth. That is also why the latter methods have lower test accuracy than the former methods in Table 1.

Latent truth learning To investigate that if LTNet can discover the latent truth given multiple inconsistent labels, we illustrate the confusion matrix between the ground truth labels and the LTNet-learned latent truth in Fig.4(b). As can be seen, the diagonal values are larger than 0.9 and most of them are larger than 0.95. The average agreement between the true labels and LTNet-learned latent truth is 0.964. Note that the LTNet was trained on the images with three sets of noisy labels. The noise percentages are 20%, 30%, and 40%, respectively.



	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	.98	.00	.01	.00	.00	.00	.00	.00	.01	.00
automobile	.00	.99	.00	.00	.00	.00	.00	.00	.00	.00
bird	.01	.00	.95	.01	.01	.01	.01	.00	.00	.00
cat	.00	.00	.01	.93	.01	.03	.01	.01	.00	.00
deer	.00	.00	.01	.01	.96	.01	.00	.01	.00	.00
dog	.00	.00	.01	.03	.01	.94	.01	.01	.00	.00
frog	.00	.00	.01	.01	.00	.00	.97	.00	.00	.00
horse	.00	.00	.00	.01	.01	.01	.00	.97	.00	.00
ship	.01	.00	.00	.00	.00	.00	.00	.00	.98	.00
truck	.00	.01	.00	.00	.00	.00	.00	.00	.00	.98

(a) (b)

Fig. 4. (a) Test accuracy curve of different methods during the training process. (b) Confusion matrix between the true labels and LTNet-learned latent truths. The LTNet is trained from the mixture of data with 20%, 30%, and 40% noises.

The average agreement between the ground truth and the three noisy labels is 0.7. If we plot confusion matrix between the true labels and the three noisy labels, the diagonal values should be about 0.8, 0.7, and 0.6 respectively. The high agreement between the ground truth and LTNet-learned latent truth indicates that LTNet is competent in discovering the latent truth from several inconsistent and noisy labels.

4.2 Evaluations on facial expression datasets

To validate the effectiveness of the proposed method in the real-world FER application, we first compared it with the state-of-the-art methods. Since errors and bias exist in the annotations of different FER datasets, we adopted a rigorous cross-dataset evaluation protocol and evaluated the methods by their average performance on 7 different datasets covering both in-the-wild and in-the-lab (posed) facial expression. Then, we analyzed the inconsistent labels in FER datasets using the proposed method.

Data Both human annotated data and unlabeled data were used in the experiments. The annotated data includes three FER datasets in-the-wild (RAF[16], AffectNet[22], and SFEW[7]) and four in-the-lab ones(CK+[20], CFEE[8], MMI[28], and Oulu-CASIA[33]).

The in-the-wild datasets contain facial expression in real world with various poses, illuminations, intensities, and other uncontrolled conditions. Both RAF and AffectNet have images downloaded from the web search engines. **RAF**[16] contains 12,271 training samples and 3,068 test samples annotated with six basic emotional categories (anger, disgust, fear, happy, sad, surprise) and neutral. Images in RAF were labeled by 315 human coders and the final annotations were determined through the crowdsourcing techniques. **AffectNet**[22] contains around 400,000 annotated images and each image is labeled by only one human coder. It includes 5,000 labeled images in 10 categories as the validation set. We selected around 280,000 images as training samples and 3,500 images as validation ones with neutral and six basic emotions. **SFEW**[7] contains images from movies annotated with neutral or one of the six basic emotions. It has 879 training samples and 406 validation samples.

The in-the-lab datasets record the facial expression in controlled environment and they usually contain posed expression. **CK+**[20] contains 593 sequences from 123 subjects, of which only 327 are annotated with 7 emotion labels (six basic emotions and contempt). We only used the ones with basic emotion labels and select the first frame of each sequence as neutral face and the last peak frame as the emotional face. Hence, 636 images were selected in total. **CFEE**[8] contains 230 subjects with 22 images each. For each subject, we selected 7 images with the six basic emotions and the neutral face. **MMI**[28] contains 30 subjects with 213 videos. For each video, we selected the first 2 images as neutral faces and the middle one third part as emotional faces. **Oulu-CASIA**[33] contains 80 subjects with 480 videos. We also selected the first 2 images as neutral faces and the last two fifth part as emotional faces.

The unlabeled data consists of the un-annotated part of AffectNet (around 700,000 images) and a collection of unlabeled facial images downloaded from Bing (around 500,000 images).

Experiment settings To evaluate the methods' generalization ability on data under the unseen condition, cross-dataset evaluation protocol was applied for SFEW, CK+, CFEE, MMI, and Oulu-CASIA datasets. In other words, only the training part of AffectNet (AffTr) and RAF (RAFTr) datasets and the unlabeled data were utilized to learn the models.

In our experiments, we adopted a 80-layer Residual Network[10] as the basic network. In the proposed IPA2LT framework, we first trained two basic models M_A and M_R from AffTr and RAFTr, respectively. Then, we used M_A to predict on RAFTr as well as the unlabeled data. Similarly, we assigned another set of annotations for AffTr and unlabeled data using M_R. The estimated annotations and the human annotations constituted the inconsistent labels, from which we trained the LTNet. Parameters in LTNet are initialized by pre-training them on the union the dataset AffTr and RAFTr. The transition layer is initialized by a close-to-identity-matrix. It is computed by adding an identity matrix and a random matrix with each entry positive. Then, each row of the initial matrix is normalized to have a sum 1. We do not initialize the probability transition matrix by the identity matrix because the identity matrix has all the non-diagonal entries as 0, which will not be updated during the training process.

The proposed LTNet was implemented under the framework of Caffe[12]. Stochastic gradient decent method was used to optimize the parameters. The momentum was 0.9 and the weight decay was 0.0005. The learning rate was initialized as 0.00001 and decreased with "poly" policy. Parameters γ and power for the learning rate policy was 0.1 and 0.5. The max iteration was set as 300,000.

Comparison with the state-of-the-art We compared the proposed method

Table 2. Test accuracy (%) of different methods on different test sets with both the in-the-wild and in-the-lab facial expressions.(**Bold**: best. **Underline**: second best.)

Test sets / Methods	in-the-wild			Posed				average		
	RAF (te.)	AffectNet (val.)	SFEW (tr+val)	CK+	CFEE	MMI	Oulu-CASIA	wild	posed	overall
AffTr (base)	79.50	56.51	55.64	91.04	**76.09**	65.32	**61.49**	63.88	73.48	69.37
RAFTr (base)	85.10	44.66	51.75	79.87	64.41	58.17	52.50	60.50	63.74	62.35
AffTr+RAFTr (base)	83.28	56.57	56.58	**92.45**	**76.09**	62.90	60.50	65.48	72.99	69.77
E2E-FC	23.99	24.00	22.33	51.73	26.52	22.25	31.28	23.44	32.95	28.87
AIR[1]	67.37	54.23	49.88	43.87	64.47	59.64	47.03	57.16	53.75	55.21
NAL[9]	84.22	55.97	58.13	91.20	75.84	64.71	61.00	66.11	73.19	70.15
IPA2LT(EM[5]+CNN)	85.30	**57.31**	54.94	86.64	72.48	63.11	59.95	65.85	70.54	68.53
IPA2LT(LTNet)	**86.77**	55.11	**58.29**	91.67	76.02	**65.61**	61.02	**66.72**	**73.58**	**70.64**

with models trained from either or both of AffTr and RAFTr, and the state-of-the-art methods addressing noisy or inconsistent labels. Table 2 presents the test accuracy of the methods on different test datasets.

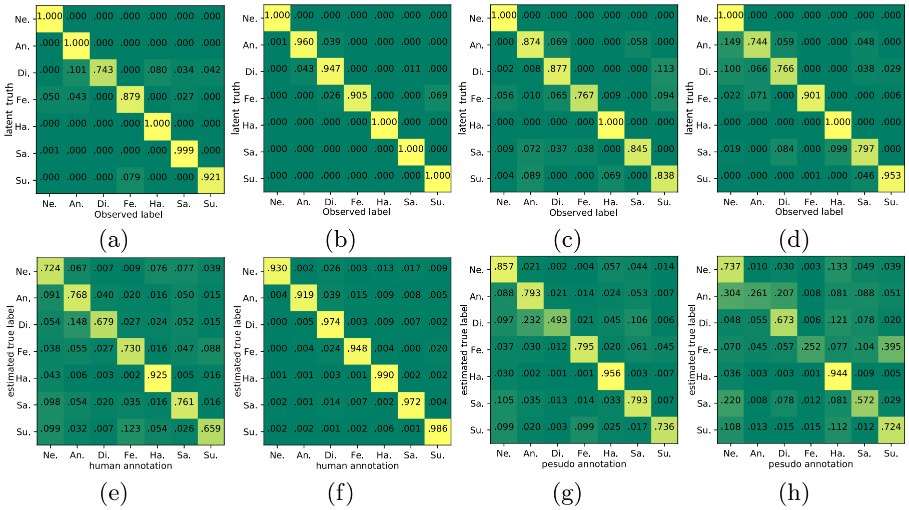
When compared to the models that are directly trained from either or both of AffTr and RAFTr, the proposed IPA2LT framework with LTNet, denoted as IPA2LT(LTNet), achieves the highest average test accuracy on in-the-wild, posed, and the overall facial expressions datasets. The consistent improvements indicate that the proposed methods cut the edge by exploring the inconsistent labels in an end-to-end training manner.

In E2E-FC, we replaced the probability transition layers in Fig. 3 with a fully connected layer that are category general but coder specific. The performance of E2E-FC is low because the probability distribution constrain is very crucial in LTNet. With the probability distribution constrain, the last second layer in LTNet can be interpreted as the hidden truth by a probability distribution. However, without the constrain, the outputs of the last second layer of E2E-FC are not essentially the reflections of the hidden truth.

AIR[1] and NAL[9] are methods that address noisy labels. In the experiments of AIR and NAL, we considered the union of AffTr and RAFTr with their human annotations as a set of noisy training data. As can be seen in Table 2, both AIR and NAL have lower test accuracy than IPA2LT(LTNet). Because AIR and NAL did not consider the annotation bias of different annotators.

We also investigated the two solutions to discover the latent truth by comparing IPA2LT(EM+CNN) and IPA2LT(LTNet). For IPA2LT(EM+CNN), we used the 2-step solution in Section 3.3 to estimate the latent truth. Results in Table 2 show that LTNet outperforms EM+CNN, because EM+CNN estimates the latent truth and trains the network separately, ignoring the relations between the input facial images and the given inconsistent labels.

Analysis of the inconsistent labels in FER To investigate whether the LTNet has learned a reasonable latent truth, we analyzed the inconsistent labels by plotting the LTNet-learned transition matrices and the confusion matrices computed from the estimated truth and the observed annotations in Fig. 5. The top row shows the transition matrices for different coders. The bottom row



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. The LTNet-learned transition matrices (top row) and the confusion matrices counted from the estimated truth and human/predicted annotations (bottom row). The top row shows transition matrices for coder (a) AffectNet, (b) RAF, (c) AffectNet-trained model, and (d) RAF-trained model. The bottom row shows statistics on dataset (e) AffectNet, (f) RAF, (g) Unlabeled data annotated by AffectNet-trained model, and (h) Unlabeled data annotated by RAF-trained model.

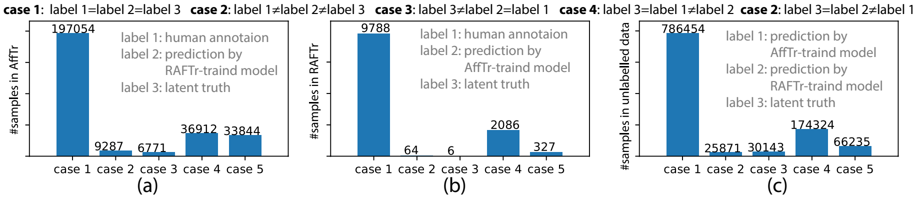


Fig. 6. Statistics of the 5 cases in (a) AffTr, (b) RAFTr, and (c) unlabeled data.

shows the confusion matrices computed from different datasets. Although the LTNet-learned transition matrices have larger diagonal values than the confusion matrices from statistics, the two rows of matrices have similar patterns. Both the transition matrix and confusion matrix with RAF dataset have the closest to 1 diagonal values. It means that the human annotations of RAF are the most reliable. That is reasonable because RAF determines a label from tens of human coders while AffectNet has only one coder each image and the unlabeled data is labeled by the trained models. We can also see from Fig. 5 (c), (d), (g), and (h) that annotations by the trained models are the least reliable.

We counted the images that have consistent and inconsistent annotations in AffTr, RAFTr, and the unlabeled data. Fig. 6 plots the statistics of samples in different cases. Case 1 contains samples with consistent human annotation,

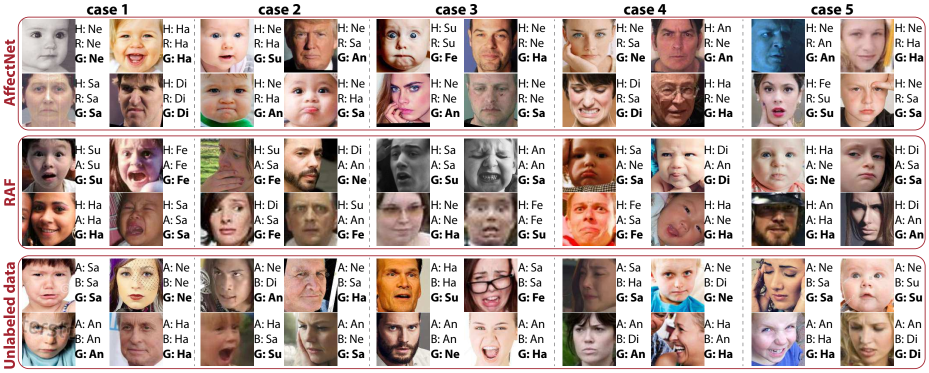


Fig. 7. Examples from the 5 cases in AffectNet, RAF, and the unlabeled data.(H: human annotation, A: prediction by AffTr-trained model, R: prediction by RAFTr-trained model, G: LTNet-learned truth. Ne: neutral, An: anger, Di: disgust, Fe: fear, Ha: happy, Sa: sad, and Su: surprise.)

latent truth, and model-predicted labels. In Case 2, all the three annotations are different from each other. In Case 3, the latent truth differs from the other two while the other two are the same. In Case 4 and 5, the latent truth agrees with one but differs from the other. Majority of the samples have consistent labels and very few of them have a latent truth that differs from both the other two labels. As can be seen from Fig. 6 (c) that the latent truth agrees more with the predictions from the AffTr-trained model, because AffTr contains much more samples than RAFTr and leads to a more robust FER model. Fig. 7 show some samples from the 5 cases in the three datasets. As can be seen, the estimated truth is reasonable whatever the other two labels are.

5 Conclusions

This paper proposed a IPA2LT framework to solve a relatively unexplored problem, i.e., how to learn a classifier from more than one datasets with inconsistent labels. To our knowledge, it is the first work to address the annotation inconsistency in different facial expression datasets. In the IPA2LT framework, we proposed an end-to-end trainable network LTNet embedded with a scheme of discovering the latent truth from multiple inconsistent labels and the input images. Experiments on both the synthetic and real data validate the effectiveness and advantages of the proposed method over other state-of-the-art methods that deal with noisy or inconsistent labels.

Acknowledgement: We gratefully acknowledge the supports from National Key R&D Program of China (grant 2017YFA0700800), National Natural Science Foundation of China (grant 61702481), and External Cooperation Program of CAS (grant GJHZ1843).

References

1. Azadi, S., Feng, J., Jegelka, S., Darrell, T.: Auxiliary image regularization for deep cnns with noisy labels. In: ICLR (2016)
2. Benitez-Quiroz, C.F., Srinivasan, R., Martinez, A.M., et al.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: CVPR. pp. 5562–5570 (2016)
3. Chen, X., Lin, Q., Zhou, D.: Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In: ICML. pp. 64–72 (2013)
4. Chu, W.S., De la Torre, F., Cohn, J.F.: Selective transfer machine for personalized facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(3), 529–545 (2017)
5. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Applied statistics pp. 20–28 (1979)
6. Dehghani, M., Severyn, A., Rothe, S., Kamps, J.: Avoiding your teacher's mistakes: Training neural networks with controlled weak supervision. arXiv preprint arXiv:1711.00313 (2017)
7. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: ICCV Workshops. pp. 2106–2112 (2011)
8. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proceedings of the National Academy of Sciences **111**(15), E1454–E1462 (2014)
9. Goldberger, J., Ben-Reuven, E.: Training deep neuralnetworks using a noise adaptation layer. In: ICLR (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
11. Hu, N., Englebienne, G., Lou, Z., Kr02se, B.: Learning to recognize human activities using soft labels. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(10), 1973–1984 (2017)
12. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
13. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Regularizing very deep neural networks on corrupted labels. arXiv preprint arXiv:1712.05055 (2017)
14. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: ICCV. pp. 2983–2991 (2015)
15. andHonglak Lee, S.E.R.: Training deep neural networks on noisy labels with bootstrapping. In: ICLR Workshops. pp. 1–11 (2015)
16. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR. pp. 2584–2593 (2017)
17. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation. In: CVPR. pp. 1910–1918 (2017)
18. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: CVPR. pp. 1805–1812 (2014)
19. Liu, Q., Peng, J., Ihler, A.T.: Variational inference for crowdsourcing. In: NIPS. pp. 692–700 (2012)
20. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: CVPR Workshops (2010)

21. Mnih, V., Hinton, G.E.: Learning to label aerial images from noisy data. In: ICML. pp. 567–574 (2012)
22. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **PP**(99), 1–1 (2017)
23. Moreno, P.G., Artés-Rodríguez, A., Teh, Y.W., Perez-Cruz, F.: Bayesian nonparametric crowdsourcing. Journal of Machine Learning Research (2015)
24. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. In: NIPS. pp. 1196–1204 (2013)
25. Pantic, M.: Facial expression recognition. In: Encyclopedia of biometrics, pp. 400–406. Springer (2009)
26. Sukhbaatar, S., Fergus, R.: Learning from noisy labels with deep neural networks. arXiv preprint arXiv:1406.2080 (2014)
27. Tian, Y., Kanade, T., Cohn, J.F.: Facial expression recognition. In: Handbook of face recognition, pp. 487–519. Springer (2011)
28. Valstar, M.F., Pantic, M.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: International Conference on Language Resources and Evaluation, Workshop on EMOTION. pp. 65–70 (2010)
29. Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S.: Learning from noisy large-scale datasets with minimal supervision. In: CVPR (2017)
30. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(1), 39–58 (2009)
31. Zhang, P., Obradovic, Z.: Learning from inconsistent and unreliable annotators by a gaussian mixture model and bayesian information criterion. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 553–568 (2011)
32. Zhang, Y., Chen, X., Zhou, D., Jordan, M.I.: Spectral methods meet em: A provably optimal algorithm for crowdsourcing. The Journal of Machine Learning Research **17**(1), 3537–3580 (2016)
33. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikäinen, M.: Facial expression recognition from near-infrared videos. Image and Vision Computing **29**(9), 607 – 619 (2011)
34. Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S.: Peak-piloted deep network for facial expression recognition. In: ECCV. pp. 425–442 (2016)
35. Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R.: Truth inference in crowdsourcing: Is the problem solved? Proc. VLDB Endow. **10**(5), 541–552 (Jan 2017)
36. Zhou, D., Liu, Q., Platt, J., Meek, C.: Aggregating ordinal labels from crowds by minimax conditional entropy. In: ICML. pp. 262–270 (2014)