Label Distribution Learning on Auxiliary Label Space Graphs for Facial Expression Recognition

Shikai Chen[1], Jianfeng Wang[2], Yuedong Chen[3], Zhongchao Shi[4], Xin Geng[*,1], Yong Rui[*,4]
[1]Southeast University [2]University of Oxford [3]Nanyang Technological University
[4]AI Lab, Lenovo Research
{skchen, xgeng}@seu.edu.cn jianfeng.wang@cs.ox.ac.uk
donald.chen@ntu.edu.sg {shizc2, yongrui}@lenovo.com

Abstract

Many existing studies reveal that annotation inconsistency widely exists among a variety of facial expression recognition (FER) datasets. The reason might be the subjectivity of human annotators and the ambiguous nature of the expression labels. One promising strategy tackling such a problem is a recently proposed learning paradigm called Label Distribution Learning (LDL), which allows multiple labels with different intensity to be linked to one expression. However, it is often impractical to directly apply label distribution learning because numerous existing datasets only contain one-hot labels rather than label distributions. To solve the problem, we propose a novel approach named Label Distribution Learning on Auxiliary Label Space Graphs(LDL-ALSG) that leverages the topological information of the labels from related but more distinct tasks, such as action unit recognition and facial landmark detection. The underlying assumption is that facial images should have similar expression distributions to their neighbours in the label space of action unit recognition and facial landmark detection. Our proposed method is evaluated on a variety of datasets and outperforms those state-of-the-art methods consistently with a huge margin.

1. Introduction

Facial Expression Recognition (FER) plays a vital role in driver assistance, health care and many other daily scenes. In recent years, the datasets of facial expression recognition have increased substantially in quantity and size [9, 22, 27, 28], significantly improving the recognition rate of some Convolutional Neural Networks (CNNs) based approaches, which incorporated facial images with optical flows [33], landmarks [40], or prior knowledge [26, 6] to enhance the performance and interpretability.

Deep learning based methods are mainly affected by three factors, which are data, model, and label [2]. Researchers have widely studied data and model but paid less heed to the label. Zeng *et al*. [39] explored the annotation subjectivity by suggesting a three-step framework trained on several inconsistently labeled datasets and abundant unlabeled data, where they did not consider the label ambi-

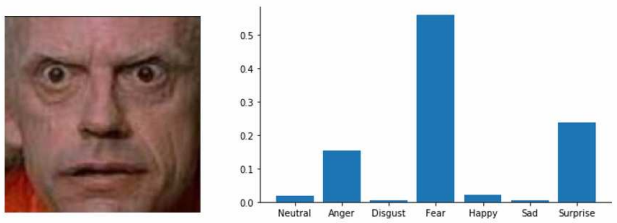

Figure 1. A real-world expressive face can be ambiguous and mixes multiple basic expressions. The label distribution on the right side is the output of the network trained in our framework.

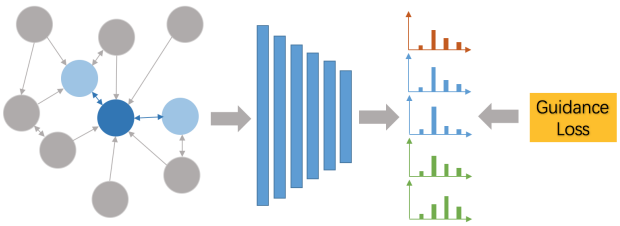

Figure 2. Overview of the proposed LDL-ALSG. With our proposed guidance loss, LDL-ALSG uses the facial images and their nearest neighbors to guide the training of the backbone network. Their neighbors are from the K-Nearest-Neighbor graphs constructed by the labels of training data in auxiliary tasks.

* X. Geng and Y. Rui are the corresponding authors.

This work was performed while Shikai Chen worked as an intern at Lenovo AI Lab.

This research was partially supported by the National Key Research & Development Plan of China (No. 2017YFB1002801), the National Science Foundation of China (61622203), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Collaborative Innovation Center of Wireless Communications Technology.

guity and the relative importance of each label. Zhou *et al*. [46] introduced an Emotion Distribution Learning(EDL) method that maps an expression image to an emotion distribution. However, Label distribution annotations needed for EDL[46] are not given by most of the facial expression recognition datasets. Xu *et al*. [36] brought forward the Graph Laplacian Label Enhancement(GLLE) to recover distribution from a logic label, which does not fit for large scale and in-the-wild datasets because of its high time complexity caused by K-Nearest-Neighbor search as well as strong assumption on feature space topology.

For real-world expressions, annotation inconsistency widely exists and can be caused by various reasons. The subjectivity of the annotation of expression labels creates bias [39], because people with different background might perceive differently. Facial expressions also incorporate a varying degree of ambiguity [46] and often combine basic expressions, especially for in-the-wild datasets (see Fig.(1)).

We addressed the annotation inconsistency by performing label distribution learning with the topological information of the auxiliary task's label space. Label distributions indicate how much each label can describe an instance, helping to handle the annotation bias and label ambiguity. Learning the relative importance of each label requires additional information covering the gap between logical labels and label distributions as most datasets do not provide label distribution annotations, making the learning difficult. Xu *et al*. [36] utilized the topological information of feature space, but their method made a strong assumption on the feature space topology. We suggested that the topological information of auxiliary task's label space guide label distribution learning and introduce more information besides current task and datasets. We assumed that facial images should have similar expression distributions to their neighbors in the label space of action unit recognition and facial landmark detection as shown in Fig.(3) and Fig.(4). Therefore, we proposed a novel approach called Label Distribution Learning on Auxiliary Label Space Graphs(LDL-ALSG) to solve the annotation inconsistency.

Our contributions are summarized as follows,

- To the best of our knowledge, the proposed LDL-ALSG is the first label distribution learning framework leveraging the label space topological information of auxiliary tasks.
- Proposed LDL-ALSG framework is end-to-end and independent of the backbone network and put no additional burden on inference.
- Proposed LDL-ALSG can effectively deal with label ambiguity and label noise, outperforming the state-of-the-art approaches with a huge margin.

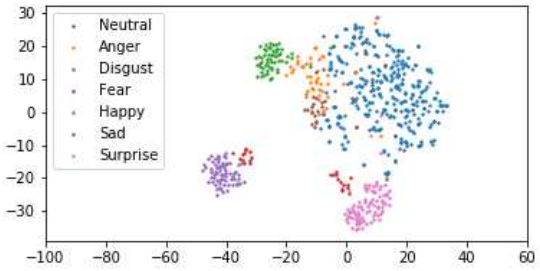

Figure 3. Visualization of action units with TSNE for CK+ dataset.

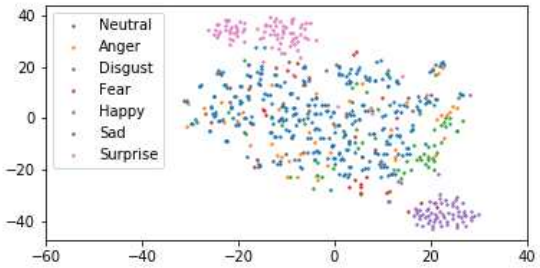

Figure 4. Visualization of facial landmarks with TSNE for CK+ dataset.

2. Related Works

2.1. Facial Expression Recognition

Facial expression recognition has remained as an active research topic during the past decades. The development of deep learning based methods has mostly centered on data and model. Traditionally, previous works have used handcrafted features to study facial expression recognition. Examples include sparse coding [45], local binary patterns (LBP) [30], non-negative matrix factorization (NMF) [44], and Gabor Wavelets [5]. Emerging deep learning-based approaches adopted to examine Facial Expression Recognition have progressed remarkably on the recognition rate for lab-controlled and in-the-wild datasets.

Deep learning-based methods require a sufficient amount of labeled training data. In the past few years, Facial Expression Recognition has been studied on numerous large-scale datasets in addition to several traditional benchmark datasets comprising CK+ [27], MMI [34], and Oulu-CASIA [42]. For instance, BU-4DFE [38] contains 60600 images of lab-controlled faces; EmotionNet [10] collected nearly one million facial expression images from the Internet; AffectNet [28] gathered up to one million facial images and manually annotated around 450,000 images; ExpW [41] downloaded about 100k in-the-wild expression images from the Internet.

Aiming to improve deep learning models, some scholars assembled network models in feature [4] and decision

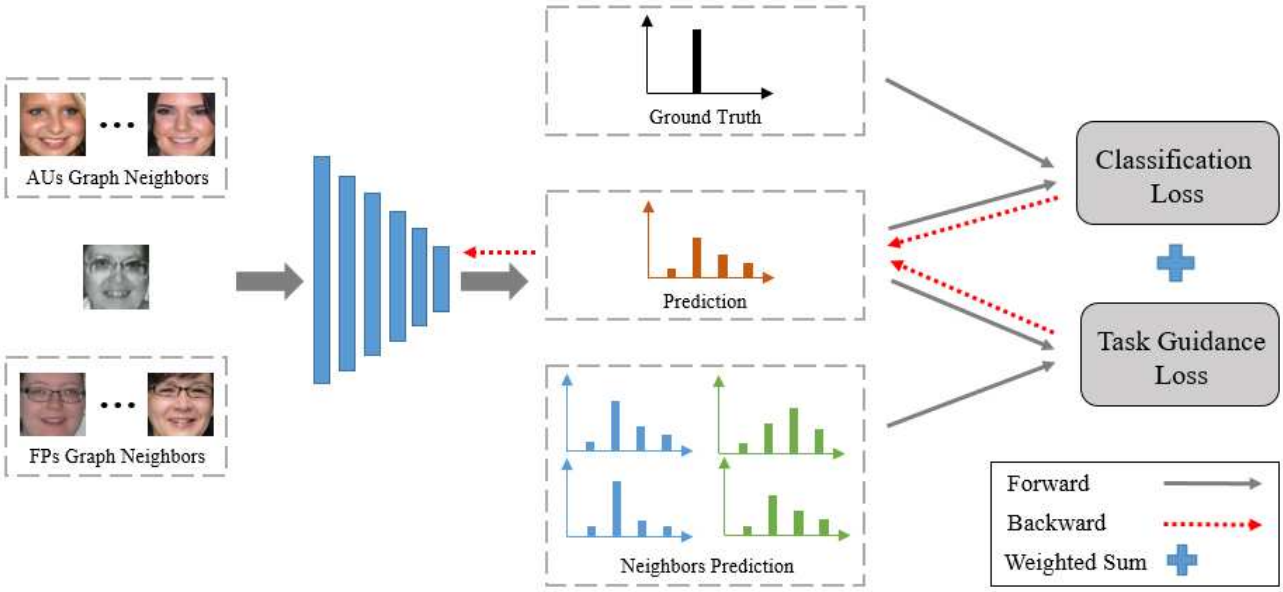

Figure 5. The framework of the proposed LDL-ALSG. Given an input facial image (denoted as the central image), its nearest neighbors can be found in the prebuilt index-similarity list for action units and facial points. Then, both the central image and its neighbors will be fed into the backbone network. The output distributions of all input images will be leveraged to minimize the classification loss and the proposed guidance loss. Because neighbor images and outputs are simply for guidance, only gradients related to the central image will be used to update the parameters of the network.

level [20] to leverage the diversity and complementarity of different structures; Some fused facial images with landmarks [45] and optical flows [33] using two stream network architectures; Others [21, 37, 6] considered the difference between the expressive face and its corresponding neutral face a valid prior knowledge.

Although extensive research has paid attention to data and model as aforementioned, only a small number of research spotlighting the label is extant.

2.2. Methods for Label Inconsistency

Early efforts primarily come from the crowdsourcing community [43, 35, 23, 7], which concentrates on estimating the ground truth out of a set of inconsistent and noisy annotations in the same dataset. Some methods leverage a small set of clean data to assess the quality of the labels during the training process [35, 23, 7]. For example, Azadi *et al*. [1] proposed AIR to train the feature extractors and Sukhbaatar *et al*. [32] suggested estimating the distribution of noisy labels. Zeng *et al*. [39] advocated the Inconsistent Pseudo Annotations to Latent Truth (IPA2LT), which can learn a classifier from more than one dataset with different annotation preferences.

These methods omit the ambiguity of facial expressions. Inconsistent annotations caused by subjectivity do not equate with incorrect annotations. We treated the inconsistent annotations as not only the noise but also labels that can describe the image for a certain degree.

2.3. Methods for Label Ambiguity and Label Noise

As Fig.(1) shows, the facial expression usually blends with different basic expressions, causing the label ambiguity and label noise that also exists in other computer vision tasks, such as head pose and facial age estimation. Label distribution learning and label enhancement have been proposed in recent years to mitigate the adverse impact of label ambiguity. As for label distribution learning, by utilizing prior knowledge, [11, 14, 13, 12, 31] transferred logical labels to discretized bivariate Gaussian label distribution, which is centered at the ground-truth label. [25, 17] established the relationship between instances and labels by graphs and transferred logical labels into label distribution. Zhou *et al*. [46] suggested an emotion distribution learning (EDL) method to deal with a more common case in which each expression associates with multiple emotion labels.

Label enhancement (LE) methods address the unavailability of label distributions. For instance, the fussy clustering based label enhancement method minimize the objective function iteratively by clustering the feature vector through C-means clustering. The kernel based label enhancement method [19] involved the kernel function in calculating the center of each feature space. The computation of the radius and the distance between samples and centers leads to membership degrees and label distributions. The label propagation based label enhancement method [24] seeks to recover the label distribution by using the iterative label

propagation technique. The manifold learning based label enhancement method [17] assumes that each data point can be reconstructed using its neighbors through a graph that represents the topological structure of the feature space. Recently, Xu *et al*. [36] came up with the Graph Laplacian Label Enhancement (GLLE) that mines the hidden importance from training instances through the topological information of the feature space. This method based on the smoothness assumption that the points close to each other are more likely to share a label.

Label enhancement methods depend on the topological information of feature space and some of them require strong assumptions such as local linear assumption and local smoothness assumption [36], remaining unfit for where features originate in deep neural networks and input data are images. Our method uses the topological information of the label space from auxiliary tasks, from which we can obtain additional information to guide the learning on the current task.

3. Label Distribution Learning on Auxiliary Label Space Graphs

The auxiliary label space is the label space of an auxiliary task that correlates to the current task and with which shares the same type of inputs, i.e., facial images. For facial expression recognition, Action Unit Recognition (AUR) and facial landmark detection (FLD), which describe facial structures and movements from different angles, are two auxiliary tasks. Besides, the label space of auxiliary tasks should be less ambiguous and can be effectively obtained through well-developed methods.

As illustrated in Fig.(2), overall, LDL-ALSG leverages the deviation between images and their neighbors' predictions to guide the training of the backbone network. A guidance loss function is proposed to utilize the label space's topological information of auxiliary tasks. Note that LDL-ALSG does not require datasets to provide distribution labels.

We selected facial landmark detection and action unit recognition as auxiliary tasks. Facial landmark detection and action unit recognition focus on facial structures and movements intimately correlated with facial expression recognition. Compared with facial expressions, which is ambiguous in nature, facial landmarks and action units are more consistent because human annotators are prone to annotate them with consistent labels. Thus, facial landmarks and action units can help to address the label inconsistency. We assumed that two nearby images in the label space of auxiliary tasks should have close label distributions with each other: if action units and facial landmarks of two facial images are nearest neighbors, their label distributions of facial expression should be similar. We visualize the CK+ dataset as an example in Fig.(3) and Fig.(4). Despite the possible unavailability of the action units and facial points, we can still benefit from many well-developed action unit recognition and facial landmark detection methods.

The main notations used in this paper are listed as follows. The instance variable is denoted by $\boldsymbol{x}$, the particular i-th instance is represented as $\boldsymbol{x}_i$, the label variable is illustrated by y, the particular j-th label value is designated as y_j , and the logical label vector of $\boldsymbol{x}_i$ is indicated by $l_i = (l^{y_1}_{\boldsymbol{x}_i}, l^{y_2}_{\boldsymbol{x}_i}, ..., l^{y_c}_{\boldsymbol{x}_i})$ in which c is the number of possible labels. The particular j-th label value of task t is denoted as y^t_j , and the label vector of $\boldsymbol{x}_i$ in auxiliary task t is indicated by $l^t_i = (l^{y^t_1}_{\boldsymbol{x}_i}, l^{y^t_2}_{\boldsymbol{x}_i}, ..., l^{y^t_k}_{\boldsymbol{x}_i})$ where k is the dimension of the label in task t. $f(\boldsymbol{x}|\boldsymbol{\theta})$ denotes the backbone model with a softmax prediction.

3.1. Problem Formulations

The process of LDL-ALSG is defined as follows: given a training set $\boldsymbol{S} = (\boldsymbol{x}_i, \boldsymbol{l}_i, \boldsymbol{l}^t_i | 1 \leq i \leq n, 1 \leq t \leq T)$, where T is the number of auxiliary tasks, LDL-ALSG is the learning process of finding a model $f(\boldsymbol{x}|\theta)$ that maps input $\boldsymbol{x_i}$ to label distribution using $\boldsymbol{l_i}$ and $\boldsymbol{l^t_i}$.

This model can be trained by solving the following problem

$$\arg\min_{\boldsymbol{\theta}} \boldsymbol{L}(\boldsymbol{\theta}) + \lambda \sum_{1}^{T} \boldsymbol{\Omega}_t(\boldsymbol{\theta}) \tag{1}$$

Where $\boldsymbol{L}$ is a loss function , $\boldsymbol{\Omega}_t(\boldsymbol{\theta})$ is the function to mine hidden label importance utilizing the topological information from the label space of related tasks.

Since the logical label can be viewed as a simplification of label distribution, we assumed that it should be close enough to the origin label to indicate the ground truth. Therefore, the L is the KL-divergence between the ground-truth label and label distribution output. Hence, we obtained the following classification loss

$$\boldsymbol{L} = \boldsymbol{KL}(l^{y_c}_{\boldsymbol{x}_i}, \boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\theta})) = \sum_{i,c} l^{y_c}_{\boldsymbol{x}_i} log(\frac{1}{f(\boldsymbol{x}_i|\boldsymbol{\theta})}), \tag{2}$$

To obtain the hidden label importance from the training data using the label space topological information of auxiliary tasks, we utilize the deviation between the network prediction of i-th central image $f(\boldsymbol{x}_i|\boldsymbol{\theta})$ and that of neighbor images j in auxiliary tasks $f(\boldsymbol{x}_j|\boldsymbol{\theta})$ to guide the update of network parameters. We need to estimate their distances or similarities to describe the relative importance of the network predictions for neighbor images. Thus, we specified the local similarity a^t_{ij} similar to GLLE[36], which is defined as

$$a^t_{ij} = \begin{cases} \exp(-\frac{||\boldsymbol{l}^t_i - \boldsymbol{l}^t_j||^2}{2\sigma^2}) & \text{if } \boldsymbol{l}^t_j \in \boldsymbol{N}^t(i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

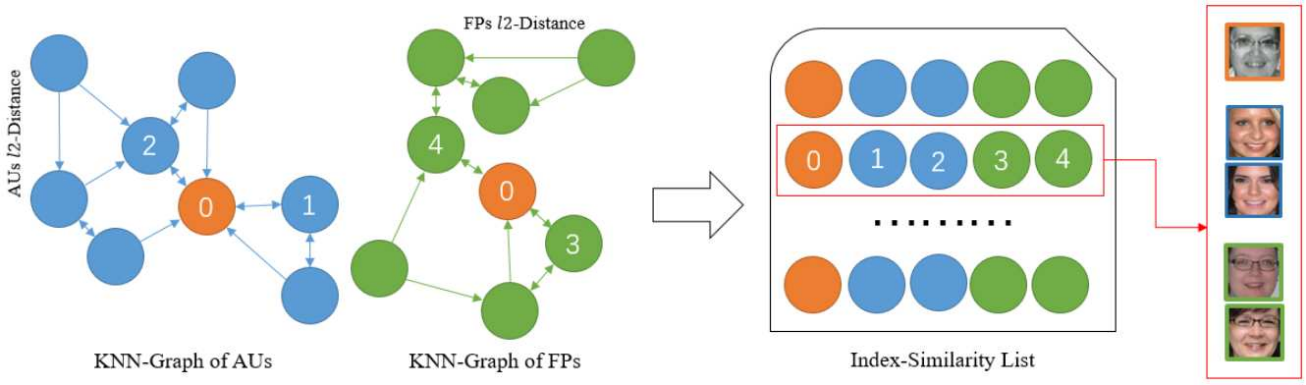

Figure 6. Approximate K-Nearest-Neighbor Graphs for the training data are built in advance. Each graph contains all the images in the training set. The indexes of each image and its neighbors will be stored in a list (denoted as the index-similarity list) along with neighbors' local similarity values. This figure is a simplified example for convenient explanation. In this example, node 0 is the central image. node 1 and node 2 are two images nearest to node 0 in action unit space. Similarly, node 3 and node 4 are two images nearest to node 0 in facial point space. Best viewed in color.

where $\boldsymbol{N}^t(i)$ means the set of $\boldsymbol{x}_i$'s K-nearest neighbors in the label space of auxiliary task t. Similar to the smoothness assumption[47], we assume that the images close to each other in the auxiliary label space are more likely to have similar expression distributions. That is to say the larger the a_{ij}^t is, the closer the distances between $f(\boldsymbol{x}_i|\boldsymbol{\theta})$ and $f(\boldsymbol{x}_j|\boldsymbol{\theta})$ could be. Hence, we have the following task guidance loss

$$\boldsymbol{\Omega}_t(\boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\theta})) = \sum_{i,j} a_{ij}^t f(\boldsymbol{x}_j|\boldsymbol{\theta}) log(\frac{f(\boldsymbol{x}_j|\boldsymbol{\theta})}{f(\boldsymbol{x}_i|\boldsymbol{\theta})}) \quad (4)$$

3.2. The overall LDL-ALSG framework

Formulating the LDL-ALSG problem into an optimization framework over Eq.(2) and Eq.(4) yields the total loss function

$$\begin{aligned} \boldsymbol{T}(\boldsymbol{\theta}) = \sum_i [\sum_c l_{\boldsymbol{x}_i}^{y_c} log(\frac{1}{f(\boldsymbol{x}_i|\boldsymbol{\theta})}) \\ + \lambda \sum_{t=1}^{T} \sum_j a_{ij}^t f(\boldsymbol{x}_j|\boldsymbol{\theta}) log(\frac{f(\boldsymbol{x}_j|\boldsymbol{\theta})}{f(\boldsymbol{x}_i|\boldsymbol{\theta})})] \end{aligned} \quad (5)$$

Our proposed LDL-ALSG is a label-side solution to expression ambiguity and annotation inconsistency so we can use any deep CNN as the backbone. After the network is fully trained, only the network parameters are needed for inference, and the trained network can independently perform prediction without any components related to auxiliary tasks.

Training data preparation: Before the training begins, the training data is prepared offline. As aforementioned, action units and facial points can be effectively extracted through well-developed methods. In our experiments, we used [3] to extract action units and facial points, which are then applied to build their own K-Nearest Neighbor (KNN) graphs separately. With the intention of bypassing the limitation of high time complexity of the kNN algorithm, we exerted an approximate KNN [18] to build the approximate kNN graphs (aKNN-graph). In our experiment, a server with 16 cpu cores can build two graphs of 300K images within 4 minutes.

For each image in the training set, we stored its index and neighbors from both graphs coupled with the local similarity values defined in Eq.(3) in a list. Then, we discarded the AKNN graphs and used the generated index-similarity list to train the backbone network, as shown in Fig.(6).

Batch generation: To effectively minimize the target function with the guidance of auxiliary tasks, each training batch consists of groups of images. Each group contains a central image together with its neighbor images and local similarity values from the generated index-similarity list as shown in Fig.(5). The local similarity values will be used to calculate guidance loss. **Forward path:** All the images in a batch will be forwarded equally, and the network will predict label distributions for each of them. With the predicted label distributions and logical ground truth labels, the task guidance loss and classification loss can be easily calculated. **Backward path:** Because we focused solely on predictions of central images, predictions of neighbor images are used to guide the update of model parameters only and do not contribute to the classification loss. Predictions of neighbor images will be detached, meaning that the total loss is back propagated only through the central images as the red lines shown in Fig.(5).

4. Experiments

4.1. Datasets

We mainly used two types of datasets, including in-the-wild datasets (RAF [22], AffectNet [28] and SFEW [8]) and

	In-the-Wild			Lab-Controlled (Posed)				Average		
	AffNet	RAF	SFEW	CK+	CFEE	MMI	Oulu-CASIA	Wild	Posed	Overall
Baseline [39]	57.97	81.81	52.19	88.99	77.22	67.79	59.35	63.99	73.34	69.33
AIR[1]	54.23	67.37	49.88	43.87	64.47	59.64	47.03	60.80	53.75	56.10
NAL[15]	55.97	84.22	**58.13**	91.20	75.84	64.71	61.00	66.11	73.19	70.15
IPA2LT (LTNet)*[39]	57.85	83.80	53.15	91.82	78.02	68.22	62.69	64.93	75.19	70.79
LDL-ALSG(AU)	58.32	85.32	56.50	91.35	77.59	**70.49**	**63.85**	**66.94**	75.85	72.01
LDL-ALSG(FL)	**59.35**	**85.53**	54.91	91.19	**78.28**	70.29	63.53	66.60	75.82	71.87
LDL-ALSG(AU+FL)	58.29	85.33	55.87	**93.08**	77.97	70.03	63.94	66.64	**76.25**	**72.13**

Table 1. Test accuracy (%) of different methods on various test sets with both in-the-wild and lab-controlled facial expressions. (**Bold**: the best. * means the results are produced by our implementation.

lab-controlled datasets (CK+ [27], CFEE [9], MMI [34] and Oulu-CASIA [42]). We reported the per-dataset based performance on the above seven datasets and the average performance on lab-controlled, in-the-wild, and all datasets.

In-the-wild datasets incorporate facial expression images in the real world. Typically, they are constructed by collecting data from the Internet and have larger scales. RAF is divided into training and test sets with a size of 12,271 and 3,068 respectively. Faces in RAF are labeled with six basic (*anger, disgust, fear, happy, sad* and *surprise*) and neutral expression. AffectNet contains more than 400k manually annotated images. We selected approximately 280,000 images as the training set and 3,500 images as the test set, all of which are labeled with six basic and neutral expressions. SFEW [8] has 879 training samples and 406 validation samples, which are collected from movies.

The lab-controlled datasets combine facial expression images recorded in the indoor controlled environment. CK+ has 593 sequences and 327 of them are annotated with seven expressions (six basic expressions and *contempt*). Each sequence starts with a neutral face and ends with a peak expression. We chose the first frame as the neutral face and the last frame as the expressive face, resulting in a total of 636 images. CFEE accommodates 230 subjects, each of which has 22 images. Only images labeled with six basic and neutral expressions are selected to conduct experiments. MMI has 213 sequences recorded from 30 subjects. Each sequence in MMI starts with a neutral face, shifts to a peak expression, and return to a neutral face in the end. In our experiments, for each sequence, the first two images are selected as neutral faces while the middle one-third part are chosen as expressive faces. Oulu-CASIA carries 480 sequences captured from 80 objects. We picked the first two images as neutral faces and the last two fifth part as expressive faces.

4.2. Experiment Settings

To make a fair comparison with other state-of-the-art approaches that focus on label-side improvement, we adopted the cross-dataset evaluation protocol to show the effectiveness of the proposed model by adhering to the settings of other methods. Specifically, we adopted only the training part of the AffectNet (AffTr) and RAF (RAFTr) datasets to train the models and use the validation set of AffectNet for validation.

Every training batch consists of 32 groups, each of which has one central image along with 4 nearest neighbors from action unit aKNN-graph and and 4 nearest neighbors from facial point aKNN-graph, totaling 9 images for one group and 288 images for each batch. After the backbone network is fully trained, we only need the backbone network and abandon the rest of the components.

Our experiments adopted the 50-layer Residual Network [16] as the backbone network, and reimplment the IPA2LT[39] method with the same backbone network and training data. The backbone network was pretrained on the training set of AffectNet and RAF. Parameters were optimized via the stochastic gradient descent method. The momentum was 0.9, the weight decay was 0.0004, and the learning rate was initialized as 0.001. Our baseline method is a directly trained resnet-50 model with cross-entropy loss and a 32 batch size. Our model and the state-of-the-art model were trained for 10 epochs with 1 epoch of linear learning rate warmup and 9 epochs of cosine learning rate decay. λ in Eq.(5) were both set to 0.0005. σ in Eq.(3) for action units and facial points were set to 1 and 68, respectively. The proposed LDL-ALSG was implemented using PyTorch[29] and trained on two Tesla V100 GPUs.

4.3. Experiment Results

Comparison with State-of-the-Art. We compared the proposed approach with existing state-of-the-art methods through extensive experiments under rigorous settings mentioned above. Table.(1) illustrates the qualtitative results. The LDL-ALSG outperforms other related approaches with a healthy margin in both lab-controlled and in-the-wild datasets.

AIR and NAL were proposed primarily to address

Training Data	Method	Accuracy		
		Wild	Lab	All
Clean data	LDL-ALSG	66.64	76.25	72.13
	IPA2LT(LTNet)	64.93	75.19	70.79
	Baseline	63.99	73.34	69.33
5% noise	LDL-ALSG	66.06	74.81	71.06
	IPA2LT(LTNet)	64.42	71.53	68.48
	Baseline	63.35	70.42	67.39
10% noise	LDL-ALSG	65.76	73.96	70.45
	IPA2LT(LTNet)		Failed	
	Baseline	61.21	70.01	66.24
15% noise	LDL-ALSG	65.03	72.40	69.24
	IPA2LT(LTNet)		Failed	
	Baseline	60.32	68.73	65.12

Table 2. Average test accuracy (%) of both baseline and LDL-ALSG methods on original training set and three synthetic datasets with label noise ratio 5%, 10% and 15%. *Clean data* refers to the mixture of training set of AffectNet and RAF datasets. *In-the-Wild* refers to the average result on the test set of AffectNet, RAF and SFEW datasets. *Lab-Controlled* refers to the average result on CK+, CFEE, MMI and Oulu-CASIA. *Overall* refers to the average results on both in-the-wild and lab-controlled datasets.

Training Data	Method	Accuracy		
		Wild	Lab	All
Clean Data	baseline	63.99	73.34	69.33
5% noise(A)	baseline	63.35	70.43	67.39
10% nosie(B)	baseline	61.21	70.01	66.24
15% noise(C)	baseline	60.32	68.73	65.12
mixture of ABC	baseline	61.92	71.34	67.30
inconsistent annotation	IPA2LT	64.01	73.56	69.46
mixture of ABC	Ours	**65.19**	**74.21**	**70.44**

Table 3. Test accuracies on datasets with inconsistent labels. Because the IPA2LT considered the annotation inconsistency explicitly, it treated the labels of the three datasets as inconsistently annotated labels for the same dataset. While our method treated the mixture of ABC as a whole dataset, the data they used was actually the same.

datasets with noisy labels. While the mixture of RAF and AffectNet datasets were taken to train the models in our settings, these two methods might not be able to deal with the annotation bias caused by different annotators, and thus failed to achieve sound performance.

LDL-ALSG and IPA2LT observed better performance than AIR and NAL, indicating that considering label inconsistency contributes to the improvement of performance. IPA2LT was not designed to cope with label ambiguity and the relative importance among different labels, so it performed worse than LDL-ALSG.

Experiment Results for Different Settings. We evaluated three different settings of LDL-ALSG and a baseline model on several benchmark datasets. LDL-ALSG(AU) used only the neighbor images from AKNN-graph of action units, and LDL-ALSG(FP) employed only the neighbor images from the AKNN-graph of facial points. LDL-ALSG(AU+FP) refers to the setting that images from both graphs are used.

As Table.(1) shows, the baseline model performs worse on all datasets than LDL-ALSG methods, which verify the adverse effect of label inconsistency. The topological information of label space from auxiliary tasks facilitates the network to learn the distribution of facial expressions and solve the problem of annotation inconsistency. The fact that LDL-ALSG(AU) slightly betters LDL-ALSG(FP) on average is not surprising because action units are more directly related to facial expressions rather than facial points.

By leveraging the topological information of the label space from both auxiliary tasks, LDL-ALSG(AU+FP) achieved the best average accuracy. Because all the methods compared here use the same validation set, higher cross-dataset performance means better generalization and less inconsistency.

Experiment Results for Label Noise and Annotation Inconsistency. To probe the influence of label noise and validate the robustness of our method, we conducted noise experiments on three synthetic datasets. Specifically, three training sets with different label noise ratio are generated by randomly revising 5%, 10%, and 15% of the corrected labels in the original training set, namely the mixture of both AffectNet and RAF's training sets. To generate label noise, we changed the ground truth label to one of the remaining six randomly. The average accuracies of in-the-wild, lab-controlled, and overall datasets are reported in Table.(2), from which we can see that the performance of proposed LDL-ALSG attained overwhelming advantages over the baseline method in all noise settings.

The IPA2LT framework is a three-step framework. In the first step, IPA2LT starts with the baseline network pre-trained on the noisy dataset and finetunes the pre-trained network on RAF and AffectNet to generate 2 coders. In the second step, auto annotations of each training set are generated by the two coders. In the last step, noisy labels and the auto annotated labels are used to finetune the pre-trained baseline network on the whole training set, which is the training part of RAF and AffectNet, using the validation set of AffectNet for validation. When RAF and AffectNet are considered as two inconsistently annotated datasets, if the noise ratio is higher than 5%, finetuning the pre-trained network cannot produce results better than that of the baseline method. Thus, the IPA2LT framework failed to handle the label noise.

Experiment Results for Annotation Inconsistency. In order to verify that our method can address the annotation inconsistency, we consider the three noisy datasets as inconsistently annotated datasets and compared our method with

Backbone Network	Method	Accuracy		
		Wild	Lab	All
Resnet-18	LDL-ALSG	65.40	73.61	70.08
	Baseline	62.37	72.35	68.07
MobilieNet V2	LDL-ALSG	65.56	75.48	71.27
	Baseline	64.71	74.01	70.03

Table 4. Average test accuracy (%) for different backbone network.

Neighbor Number		λ	Accuracy		
AU	FP		Wild	Lab	All
2	0	0.001	66.81	**75.86**	71.98
4	0	0.001	**66.94**	75.82	**72.01**
8	0	0.001	66.69	75.44	71.69
0	2	0.001	66.30	75.43	71.52
0	4	0.001	**66.60**	**75.82**	**71.87**
0	8	0.001	66.59	75.64	71.76

Table 5. Comparison of different neighbor numbers. The λ was fixed to 0.001 and the number of neighbors in action unit space and facial point space were evaluated seperatly.

the state-of-the-art method. As shown in Table.(3), both of our method and the IPA2LT outperformed the baseline method. Our method was directly trained with the mixed datasets, while the IPA2LT framework followed its three-step training process. The proposed LDL-ALSG produced the best results showing that the topological information of auxiliary task's label space can help address the annotation inconsistency.

4.4. Ablation Study

Since hyper-parameters λ in Eq.(5) balance the classification loss and the guidance losses and the number of nearest neighbor is vital to the overall performance as well, we evaluated their influences in different settings.

We set λ to 0.001 and evaluated the influence of the number of nearest neighbors. As shown in Table.(5), different number of neighbors produced similar results but using 4 neighbors is the best. The reason might be that using too many neighbors might include neighbors that are far away from the central image in terms of auxiliary label space distances, while using too few neighbors cannot learn the relative importance of each label from limited neighbors. We also fixed neighbor numbers to 4 and study the influence of λ for guidance loss of action units and facial points. Results in Table.(6) showed that all of the different settings can outperform the baseline result and setting λ to 0.001 for both of the auxiliary tasks produced the better results.

When the nearest neighbors from both graphs were utilized to guide the learning process, the influence of the nearest neighbor number and the λ was evaluated, resulting in

Neighbor Number		λ	Accuracy		
AU	FP		Wild	Lab	All
4	0	0.0001	66.36	75.58	71.61
4	0	0.001	**66.94**	**75.82**	**72.01**
4	0	0.01	66.33	74.96	71.27
0	4	0.0001	66.54	75.78	71.82
0	4	0.001	**66.60**	**75.82**	**71.87**
0	4	0.01	65.89	75.45	71.36

Table 6. Comparison of different λ. The number of neignbors was fixed to 4 and the λ of neighbors in action unit space and facial point space were evaluated seperatly.

Neighbor Number		λ	Accuracy		
AU	FP		Wild	Lab	All
4	4	0.0001	**67.13**	75.84	72.11
4	4	0.0005	66.64	**76.25**	**72.13**
4	4	0.001	66.42	75.52	71.62
4	4	0.01	66.34	75.68	71.68

Table 7. Comparison of different λ, when utilizing topological information from both graphs. The number of neighbors was fixed to 4.

Table.(7). When guidance loss from both tasks were combined, setting λ to 0.0005 produced the best results.

We also performed experiments on two different backbone network to show that our method is network independent as shown in Table.(4).

5. Conclusion

In this paper, we proposed the LDL-ALSG framework to tackle annotation inconsistency and label ambiguity in facial expression recognition datasets. The label space topological information of auxiliary tasks such as action unit recognition and facial landmark detection is used to help in learning the ambiguous and subjective labels. It helps to address the label inconsistency in facial expression recognition datasets because human annotators are more likely to annotate facial landmarks and action units with consistent labels. We treated the inconsistent annotations as not only the noise but also labels that can describe the data for a certain degree and train the network to predict label distributions.

The LDL-ALSG framework is an end-to-end label-side solution that is network-independent with no additional components needed for inference after the backbone network finishes training. To our knowledge, the LDL-ALSG framework is the first work that performs label distribution learning leveraging label space topological information. Our method does not need to generate label distributions in advance. Experiments on several datasets validated the effectiveness and advantages of the proposed method.

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] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[4] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 433–436. ACM, 2016.

[5] Juliano J Bazzo and Marcus V Lamar. Recognizing facial actions using gabor wavelets with neutral face average difference. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 505–510. IEEE, 2004.

[6] Yuedong Chen, Jianfeng Wang, Shikai Chen, Zhongchao Shi, and Jianfei Cai. Facial motion prior networks for facial expression recognition. *arXiv preprint arXiv:1902.08788*, 2019.

[7] Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. Avoiding your teacher's mistakes: Training neural networks with controlled weak supervision. *arXiv preprint arXiv:1711.00313*, 2017.

[8] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, 2011.

[9] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[10] 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.

[11] B. B. Gao, C. Xing, C. W. Xie, J. Wu, and X. Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, PP(99):1–1, 2016.

[12] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[13] Xin Geng and Yu Xia. Head pose estimation based on multivariate label distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1837–1842, 2014.

[14] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013.

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

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Peng Hou, Xin Geng, and Min-Ling Zhang. Multi-label manifold learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[18] Masajiro Iwasaki and Daisuke Miyazaki. Optimization of indexing based on k-nearest neighbor graph for proximity search in high-dimensional data. *arXiv preprint arXiv:1810.07355*, 2018.

[19] Xiufeng Jiang, Zhang Yi, and Jian Cheng Lv. Fuzzy svm with a new fuzzy membership function. *Neural Computing & Applications*, 15(3-4):268–276, 2006.

[20] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013.

[21] Youngsung Kim, ByungIn Yoo, Youngjun Kwak, Changkyu Choi, and Junmo Kim. Deep generative-contrastive networks for facial expression recognition. *arXiv preprint arXiv:1703.07140*, 2017.

[22] 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.

[23] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017.

[24] Yu-Kun Li, Min-Ling Zhang, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *2015 IEEE International Conference on Data Mining*, pages 251–260. IEEE, 2015.

[25] Yu Kun Li, Min Ling Zhang, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *IEEE International Conference on Data Mining*, 2016.

[26] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *FG*, pages 1–6, 2013.

[27] 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.

[28] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, va-

lence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[30] 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.

[31] Kai Su and Xin Geng. Soft facial landmark detection by label distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5008–5015, 2019.

[32] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *Eprint Arxiv*, 2014.

[33] Ning Sun, Qi Li, Ruizhi Huan, Jixin Liu, and Guang Han. Deep spatial-temporal feature fusion for facial expression recognition in static images. *Pattern Recognition Letters*, 2017.

[34] 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, 2010.

[35] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 839–847, 2017.

[36] Ning Xu, An Tao, and Xin Geng. Label enhancement for label distribution learning. In *IJCAI*, pages 2926–2932, 2018.

[37] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[38] L Yin, X Chenand Y Sun, T Worm, and M Reale. A high-resolution 3d dynamic facial expression database, 2008. In *IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands*, volume 126, 2008.

[39] 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.

[40] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017.

[41] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.

[42] 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.

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

[44] Ruicong Zhi, Markus Flierl, Qiuqi Ruan, and W Bastiaan Kleijn. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(1):38–52, 2011.

[45] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2562–2569. IEEE, 2012.

[46] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1247–1250. ACM, 2015.

[47] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, language technologies institute, school of . . . , 2005.