Concentration Distribution Learning from Label Distributions

ICML 2025

Background

- **Single-Label Learning(SLL):**

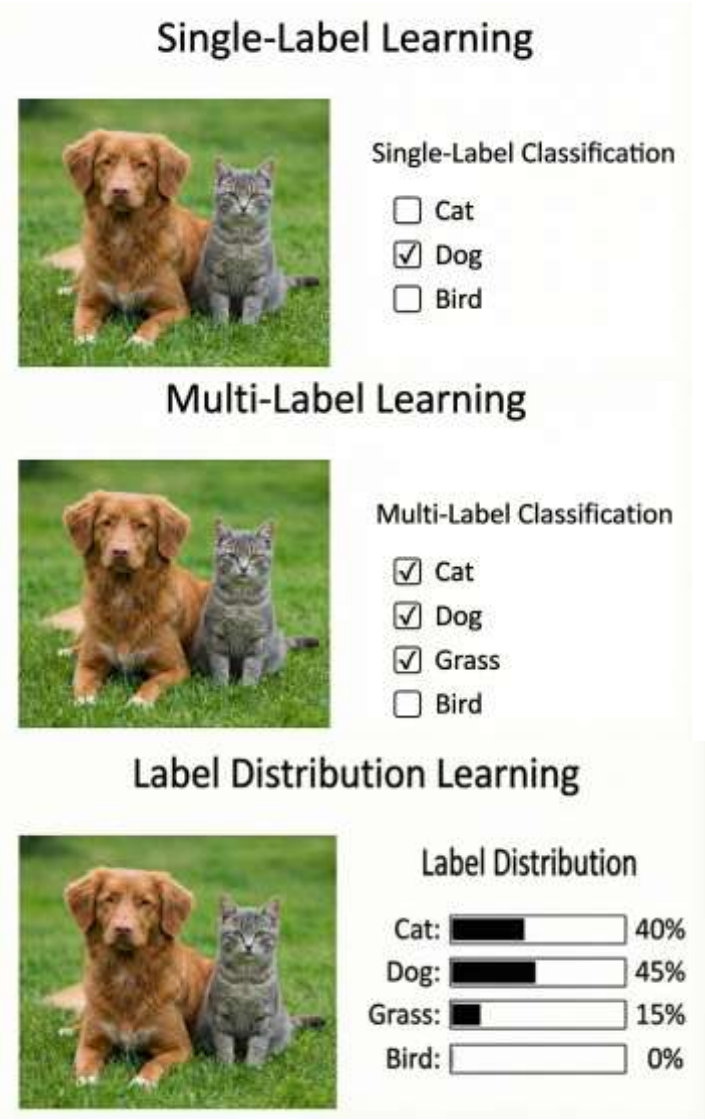
One instance corresponds to only one label.

- **Multi-Label Learning(MLL):**

One instance corresponds to a set of related labels.
Use logical value(0 or 1) to indicate whether a label is capable of describing a sample.
Can't accurately describe each label's relative importance to a sample.

- **Label-Distribution Learning(LDL):**

Use real numbers to illustrate how important a label is to a particular sample. Let d_x^l stand for the description degree of the label l to the instance x, and it's governed by:
$d_x^l \in [0,1]$ and $\sum_l d_x^l = 1$

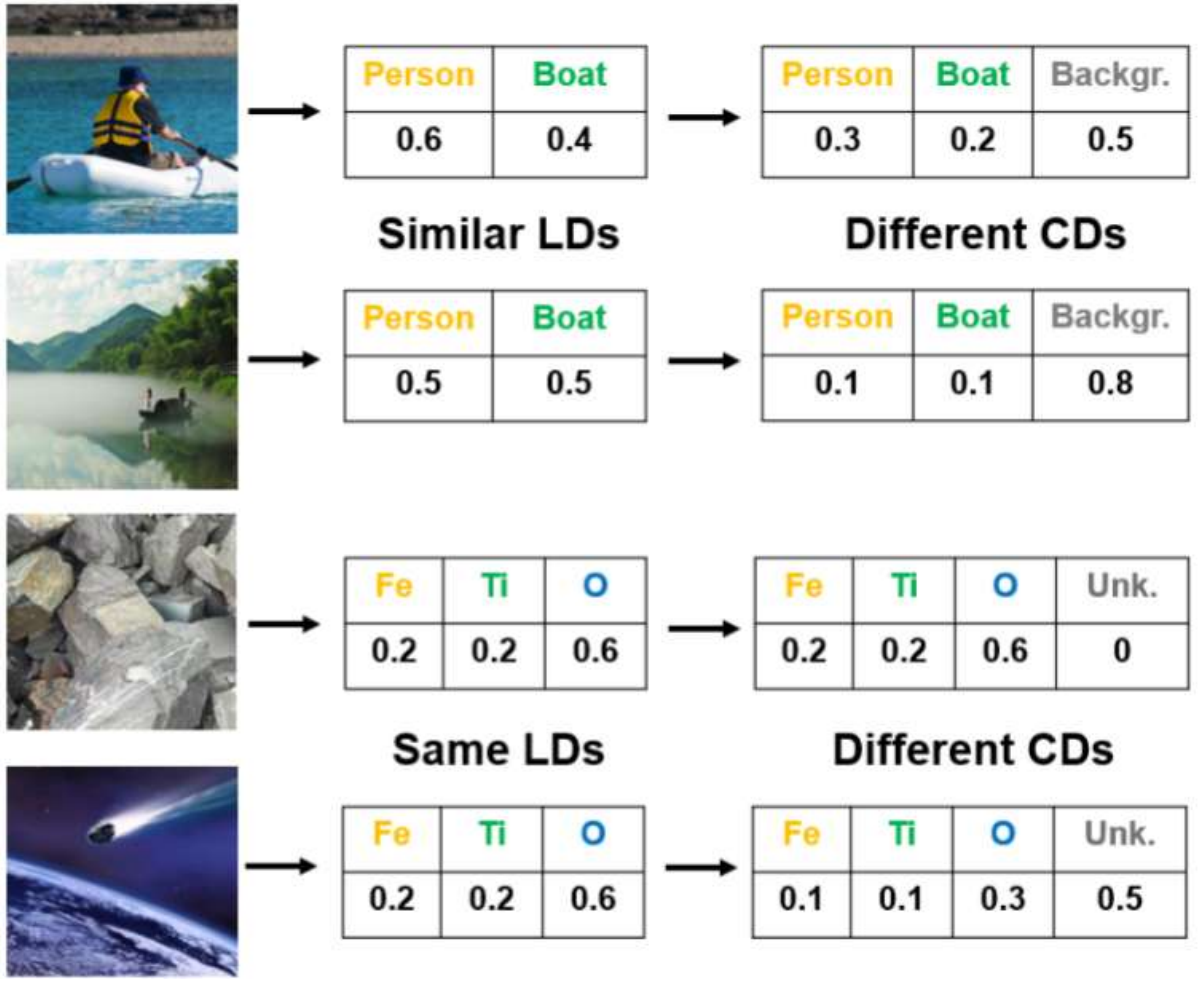


Background

- LDL only models '**relative concentration**' while ignoring '**absolute concentration**'.
- Ground-truth LDs are **incomplete representation,** cannot describe the '**hidden labels**' or '**background information**' (or **noise**) that are not defined in the label space.

How to solve it?

- Introduce the concept of **background concentration**, which represents the description degree of the complementary set of all existing labels.
- Appending background concentration terms to the original LDs, we get **concentration distributions (CDs).**



Background

This paper defines **concentration distribution learning (CDL)** as the process of learning the background concentration μ and the label distribution b simultaneously,

Learn concentration distribution vectors $\boldsymbol{c_d} = [\boldsymbol{b}, \mu]$ from concentration distribution datasets, and meet the constraints:

$$\mu > 0, \forall i \in [1, 2, ..., c], b_i \in [0, 1] \qquad \sum_{i=1}^{c} b_i + \mu = 1.$$

Method

Key Assumptions : The distribution of labels observed in the dataset is called the **apparent label distribution**.

The relationship between the apparent label distribution vector and the concentration distribution vector can be formulated as :

$$\boldsymbol{p} = \boldsymbol{b} + \boldsymbol{\mu}^*,$$

$$\boldsymbol{\mu}^* \in \mathbb{R}^c, \forall i \in [1, 2, ..., c], \boldsymbol{\mu}^*{}_i \in [0, 1] \text{ and } \sum_{i=1}^{c} \boldsymbol{\mu}^*{}_i = \mu.$$

The distribution of the background concentration μ on the real label distribution vector *b* converts the concentration distributions to the apparent label distributions. In other words, the apparent label distribution is affected by both the dataset and the background concentration.

How to solve for *b* and μ?

Assume that p obeys the Dirichlet distribution, i.e., $\boldsymbol{p} \sim Dir(\boldsymbol{\alpha})$

The expectation on p_i can be written as

$$\mathbb{E}_{p_i} = \frac{\alpha_i}{\sum_{j=1}^{c} \alpha_j}. \qquad \alpha_i = e_i + u_i,$$

- α_i : the belief mass on the i-th class
- e_i : dataset-side confidence
- u_i : background confidence

Method

Boundary conditions: **when there is nothing in the dataset to provide confidence**

$$\forall i \in [1, 2, ..., c], \boldsymbol{e} = [e_1, e_2, ..., e_c] = \mathbf{0}_c,$$

$$\mathbf{1}_c = \mathbf{0}_c + \boldsymbol{u}$$

$$\boldsymbol{u} = \mathbf{1}_c,$$

$$\alpha_i = e_i + 1.$$

$$\mathbb{E}_{p_i} = \frac{e_i + 1}{\sum_{j=1}^{c}(e_j + 1)} = \frac{e_i + 1}{\sum_{j=1}^{c} e_j + c}.$$

Assuming that the background concentration is evenly spread on each class of the real label distribution vector *b* in probability

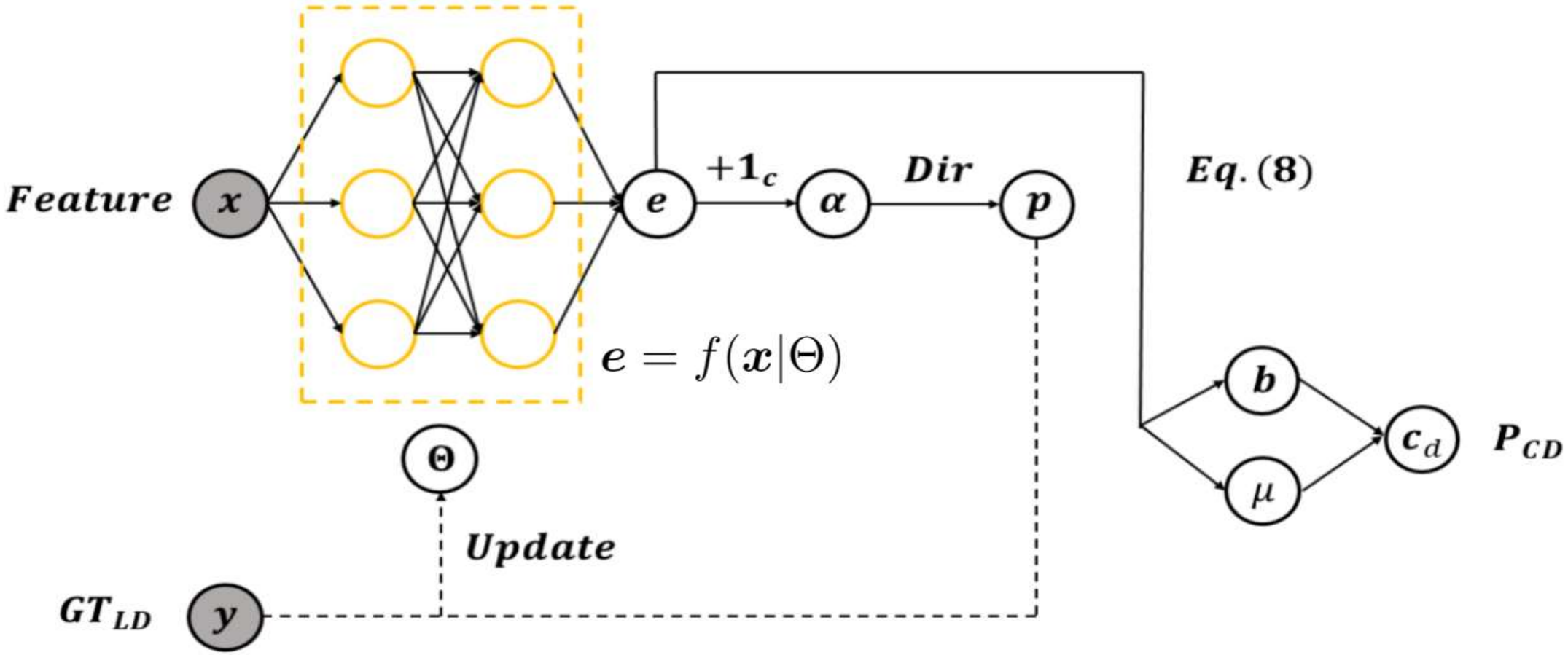
$$\forall i \in [1, 2, ..., c], \mathbb{E}_{\boldsymbol{\mu}_i^*} = \frac{\mu}{c},$$

$$\mathbb{E}_{p_i} = b_i + \frac{\mu}{c}.$$

$$b_i = \frac{e_i}{\sum_{j=1}^{c} e_j + c}, \mu = \frac{c}{\sum_{j=1}^{c} e_j + c}$$

Method

Framework of method



The softmax layer of a conventional neural network is replaced with an activation function layer to ensure that the network outputs non-negative values

Method

In conventional neural network:

$$\mathcal{L}_{MSE} = ||\boldsymbol{y} - \boldsymbol{p}||_2^2,$$

In this model:

$$\begin{aligned}\mathcal{L}_{AMSE}(\boldsymbol{\alpha}) &= \int ||\boldsymbol{y} - \boldsymbol{p}||_2^2 \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{c} p_i^{\alpha_i - 1} d\boldsymbol{p} \\ &= \sum_{i=1}^{c} \underbrace{(y_i - \frac{\alpha_i}{S})^2}_{\mathcal{L}_{err}} + \underbrace{\frac{\alpha_i(S-\alpha_i)}{S^2(S+1)}}_{\mathcal{L}_{var}} \\ &= \sum_{i=1}^{c} (y_i - \hat{p}_i)^2 + \frac{\hat{p}_i(1-\hat{p}_i)}{S+1},\end{aligned}$$

where $S = \sum_{i=1}^{c} \alpha_i$ and $\hat{p}_i = \frac{\alpha_i}{S}$

Method

Generalization Bound

Definition 1. *Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to [0,1] and $\mathcal{S}$ be a set of fixed samples with size n. Then, the empirical Rademacher complexity of $\mathcal{H}$ with respective to $\mathcal{S}$ is defined as*

$$\begin{aligned}\widehat{\mathcal{R}}_S(\mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h\left(x_i\right)\right] \\ &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathcal{L}_{AMSE}\left(\alpha_i\right)\right],\end{aligned} \tag{11}$$

Lemma 1. *Let $\mathcal{H}$ be a family of functions. For any $\delta > 0$, with probability at least $1-\delta$, for all $h \in \mathcal{H}$ such that*

$$\mathcal{L}(h) \leq \mathcal{L}_S(h) + \widehat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2n}}, \tag{12}$$

$$\begin{aligned}\mathcal{L}(h) \leq& \mathcal{L}_S(h) + \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathcal{L}_{AMSE}\left(\alpha_i\right)\right] \\ &+ 3\sqrt{\frac{\log 2/\delta}{2n}}.\end{aligned} \tag{13}$$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathcal{L}_{AMSE}\left(\alpha_i\right)\right] &\leq \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{AMSE}\left(\alpha_i\right) \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\left[(y_{ij}-\frac{\alpha_{ij}}{S_i})^2 + \frac{\alpha_{ij}(S_i-\alpha_{ij})}{S_i^2(S_i+1)}\right] \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\left[(y_{ij}-\frac{\alpha_{ij}}{S_i})^2 + \frac{\frac{\alpha_{ij}}{S_i}(1-\frac{\alpha_{ij}}{S_i})}{S_i(S_i+1)}\right] \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left[1+\frac{1}{4c(c+1)}\right]\end{aligned} \tag{14}$$

$$\mathcal{L}(h) \leq \mathcal{L}_S(h) + \frac{1}{n}\sum_{i=1}^{n}\left[1+\frac{1}{4c(c+1)}\right] + 3\sqrt{\frac{\log 2/\delta}{2n}}. \tag{15}$$

$$\mathcal{L}(h) \leq \mathcal{L}_S(h) + \underbrace{\left[1+\frac{1}{4c(c+1)}\right]}_{bound}. \tag{16}$$

Construction of CDL dataset

SJAFFE Dataset --> SJA_c

Original: $\boldsymbol{s} = [3, 4.8, 1.2, 2.1, 2.4, 1.5]$

$$\boldsymbol{s}_n = \frac{[3, 4.8, 1.2, 2.1, 2.4, 1.5]}{3 + 4.8 + 1.2 + 2.1 + 2.4 + 1.5} = [0.2, 0.32, 0.08, 0.14, 0.16, 0.1].$$

Background concentration: $\mu = 5 * 6 - \sum_{i=1}^{6} \boldsymbol{s}_i,$

Final:

$$\boldsymbol{c}_d = \frac{[3, 4.8, 1.2, 2.1, 2.4, 1.5, 15]}{3 + 4.8 + 1.2 + 2.1 + 2.4 + 1.5 + 15} = [0.1, 0.16, 0.04, 0.07, 0.08, 0.05, 0.5].$$

Experiments Setting up

For each dataset, the last class of the ground-truth label distribution is regarded as the background concentration and will be invisible in training.

For baseline methods, append $g + \delta$ as the predicted background concentration to the predicted label distribution vector, where g is the ground-truth description degree of the last class and $-0.2g < \delta < 0.2g$ is a random noise to simulate the inaccurate learning of background concentration in the baseline methods.

Experiment

	Metric	CDL-LD	LDLLC	SA-IIS	LCLR	LDLFs	LDLLDM	DLDL
Alp	Cheby	**0.0245±.0002(1)**	0.0571±.0022(7)	0.0366±.0028(2)	0.0529±.0018(6)	0.0502±.0005(3)	0.0523±.0013(5)	0.0520±.0002(4)
	Clark	**0.2797±.0009(1)**	0.4586±.0128(6)	0.7340±.0297(7)	0.4410±.0002(5)	0.4253±.0066(2)	0.4365±.0211(4)	0.4318±.0049(3)
	KL	**0.0098±.0000(1)**	0.0914±.0018(7)	0.0623±.0487(4)	0.0858±.0020(6)	0.0835±.0005(5)	0.0365±.0028(3)	0.0351±.0001(2)
	Cosine	**0.9894±.0017(1)**	0.9560±.0019(7)	0.9824±.0027(2)	0.9563±.0021(6)	0.9590±.0007(3)	0.9570±.0009(5)	0.9573±.0001(4)
Cdc	Cheby	**0.0291±.0002(1)**	0.0619±.0068(4)	0.1456±.0119(7)	0.0565±.0022(2)	0.0632±.0003(6)	0.0599±.0008(3)	0.0626±.0037(5)
	Clark	**0.2891±.0003(1)**	0.4876±.0270(6)	0.8329±.0030(7)	0.4327±.0007(2)	0.4539±.0026(4)	0.4457±.0122(3)	0.4611±.0086(5)
	KL	**0.0122±.0003(1)**	0.1114±.0082(6)	0.1327±.0506(7)	0.0959±.0010(4)	0.1031±.0007(5)	0.0426±.0041(2)	0.0448±.0041(3)
	Cosine	**0.9870±.0005(1)**	0.9485±.0079(5)	0.8489±.0164(7)	0.9565±.0024(2)	0.9479±.0003(6)	0.9530±.0007(3)	0.9502±.0047(4)
Col	Cheby	**0.0821±.0009(1)**	0.1606±.0077(7)	0.1561±.0312(6)	0.1475±.0084(4)	0.1488±.0030(5)	0.1441±.0013(2)	0.1454±.0060(3)
	Clark	**0.2073±.0002(1)**	0.4062±.0181(6)	0.4670±.1468(7)	0.3734±.0218(4)	0.3849±.0015(5)	0.3674±.0064(2)	0.3719±.0146(3)
	KL	**0.0262±.0006(1)**	0.2421±.0135(7)	0.1051±.0599(4)	0.2180±.0151(5)	0.2209±.0044(6)	0.0973±.0064(2)	0.0985±.0100(3)
	Cosine	**0.9737±.0006(1)**	0.9139±.0058(7)	0.9385±.0219(2)	0.9266±.0056(6)	0.9277±.0022(5)	0.9302±.0015(3)	0.9289±.0044(4)
Dia	Cheby	**0.0677±.0004(1)**	0.1092±.0014(4)	0.1529±.0755(7)	0.1140±.0004(6)	0.1105±.0100(5)	0.1079±.0033(3)	0.1072±.0111(2)
	Clark	**0.2910±.0001(1)**	0.4420±.0055(6)	0.5501±.0090(7)	0.4394±.0006(5)	0.4308±.0219(4)	0.4186±.0058(2)	0.4227±.0188(3)
	KL	**0.0276±.0003(1)**	0.1707±.0046(6)	0.0961±.0236(4)	0.1721±.0010(7)	0.1689±.0114(5)	0.0756±.0010(3)	0.0732±.0098(2)
	Cosine	**0.9718±.0003(1)**	0.9292±.0001(5)	0.9042±.0677(7)	0.9259±.0004(6)	0.9294±.0102(4)	0.9315±.0028(2)	0.9344±.0105(3)
SJA	Cheby	0.4335±.0002(7)	0.3484±.0473(3)	0.3686±.0326(4)	**0.3045±.0056(1)**	0.3734±.0119(5)	0.3388±.0219(2)	0.3790±.0216(6)
	Clark	**0.9992±.0002(1)**	1.1948±.0591(7)	1.0976±.0239(4)	1.0627±.0053(3)	1.0373±.0722(2)	1.1630±.0126(6)	1.1508±.0736(5)
	KL	**0.4841±.0007(1)**	1.5855±.0465(7)	0.7308±.0122(2)	0.7653±.0718(3)	0.7871±.1415(4)	1.1815±.0159(6)	0.9323±.0807(5)
	Cosine	**0.8161±.0013(1)**	0.7211±.0425(6)	0.7038±.0316(7)	0.7478±.0021(4)	0.7587±.0110(3)	0.7222±.0149(5)	0.7597±.0398(2)
SBU	Cheby	0.3362±.0008(3)	0.3688±.0047(5)	0.3511±.0213(4)	**0.2972±.0019(1)**	0.3123±.0019(2)	0.3723±.0151(6)	0.3823±.0171(7)
	Clark	**0.9682±.0035(1)**	1.1914±.0079(7)	1.1503±.0263(5)	0.9798±.0207(3)	0.9757±.0024(2)	1.1761±.0418(6)	1.1245±.0010(4)
	KL	**0.4641±.0028(1)**	1.4267±.0065(7)	0.7209±.0103(2)	0.7894±.0360(3)	0.8296±.0114(4)	1.2469±.0190(5)	1.2848±.0215(6)
	Cosine	0.6310±.0009(7)	0.7060±.0044(4)	0.7183±.0208(3)	**0.7578±.0008(1)**	0.7499±.0013(2)	0.7040±.0097(5)	0.6922±.0255(6)
Sce	Cheby	**0.3948±.0011(1)**	0.5289±.0033(4)	0.5438±.0447(6)	0.4789±.0021(3)	0.4765±.0083(2)	0.5363±.0154(5)	0.5440±.0153(7)
	Clark	**1.9714±.0041(1)**	2.4907±.0019(2)	2.7042±.0292(7)	2.5404±.0004(6)	2.5387±.0037(5)	2.5202±.0065(3)	2.5237±.0333(4)
	KL	**0.4918±.0010(1)**	1.1354±.0108(2)	1.7242±.0340(7)	1.5624±.0077(5)	1.5938±.0605(6)	1.2310±.1056(3)	1.2955±.0116(4)
	Cosine	**0.6063±.0020(1)**	0.5787±.0048(2)	0.4772±.0116(6)	0.5282±.0011(3)	0.4629±.0083(7)	0.5187±.0072(5)	0.5198±.0234(4)
SJAc	Cheby	**0.1153±.0031(1)**	0.3190±.0096(4)	0.3024±.0014(3)	0.2504±.0003(2)	0.3598±.0069(7)	0.3303±.0148(5)	0.3518±.0065(6)
	Clark	**0.5336±.0048(1)**	0.9786±.0076(5)	0.8086±.0100(3)	0.7821±.0076(2)	1.1820±.0341(7)	0.9712±.0124(4)	0.9983±.0164(6)
	KL	**0.0722±.0020(1)**	0.7183±.0020(6)	0.2021±.0088(2)	0.4076±.0017(3)	1.0728±.0394(7)	0.4573±.0346(4)	0.5089±.0277(5)
	Cosine	**0.9740±.0062(1)**	0.7459±.0194(4)	0.8824±.0192(2)	0.8434±.0046(3)	0.7160±.0489(6)	0.7195±.0154(5)	0.6791±.0072(7)

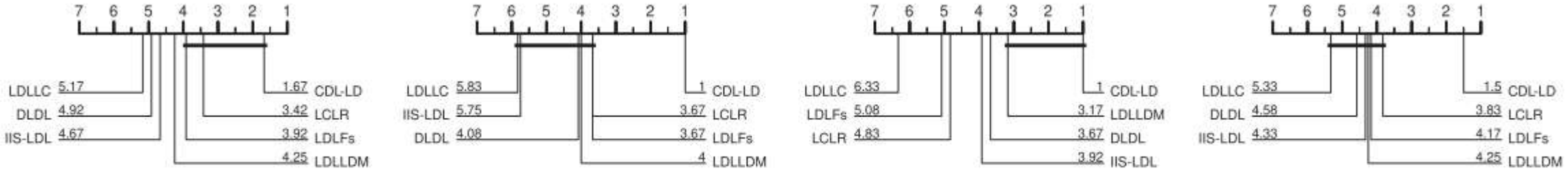
Average ranks in terms of four metrics

	CDL-LD	LDLLC	SA-IIS	LCLR	LDLFs	LDLLDM	DLDL
Cheby	**1.67(1)**	5.17(7)	4.67(5)	3.42(2)	3.92(3)	4.25(4)	4.92(6)
Clark	**1.00(1)**	5.83(7)	5.75(6)	3.67(2)	3.67(2)	4.00(4)	4.08(5)
KL	**1.00(1)**	6.33(7)	3.92(4)	4.83(5)	5.08(6)	3.17(2)	3.67(3)
Cosine	**1.50(1)**	5.33(7)	4.33(5)	3.83(2)	4.17(3)	4.25(4)	4.58(6)

Average ranks in terms of running time

	CDL-LD	LDLLC	SA-IIS	LCLR	LDLFs	LDLLDM	DLDL
Alp	**10.45(1)**	177.12(6)	69.34(2)	78.75(3)	136.89(5)	114.36(4)	180.71(7)
SBU	**29.96(1)**	189.84(5)	103.29(2)	123.57(3)	158.09(4)	197.38(6)	241.15(7)
Sce	**17.23(1)**	186.58(6)	92.62(3)	92.03(2)	143.62(4)	172.51(5)	234.02(7)

Significance Tests



(a) Chebyshev (b) Clark (c) KL (d) Cosine

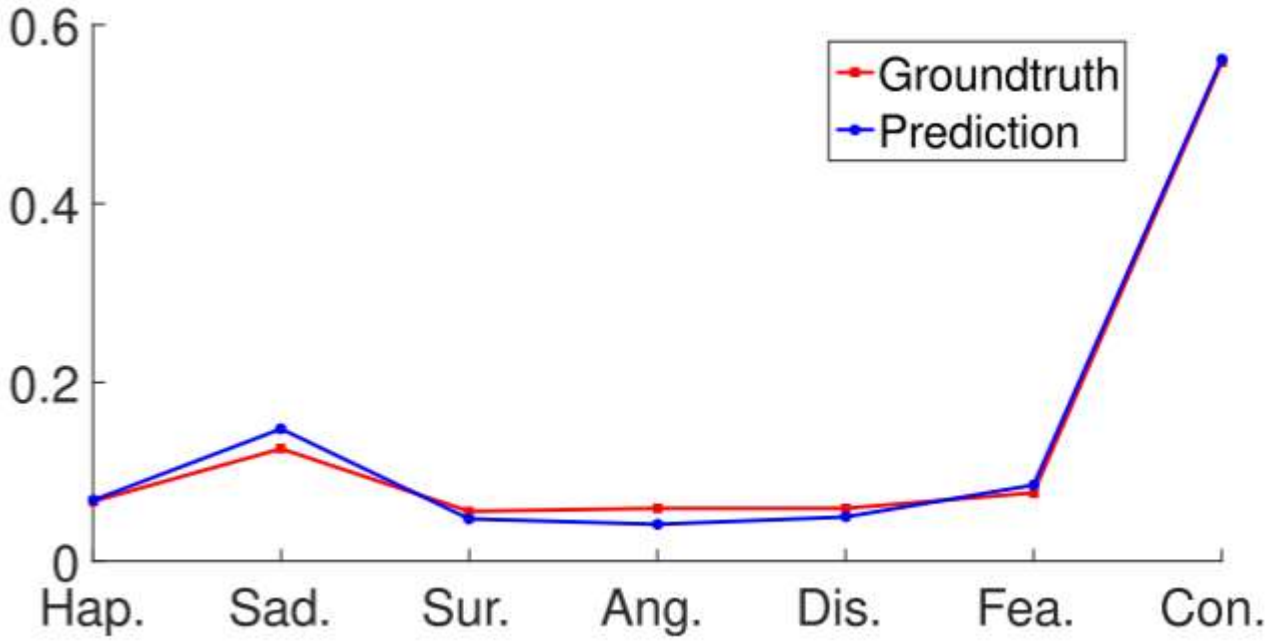
Bonferroni-Dunn test (CD = 2.3265 at 0.05 significance level)

Critical Value	Evaluation metric	Chebyshev	Clark	KL	Cosine
2.913	Friedman Statistics F_F	22.2645	40.2058	43.9416	21.8687

Friedman statistics F_F with the critical value at a significance level of 0.05

Visualization

In SJA_c dataset



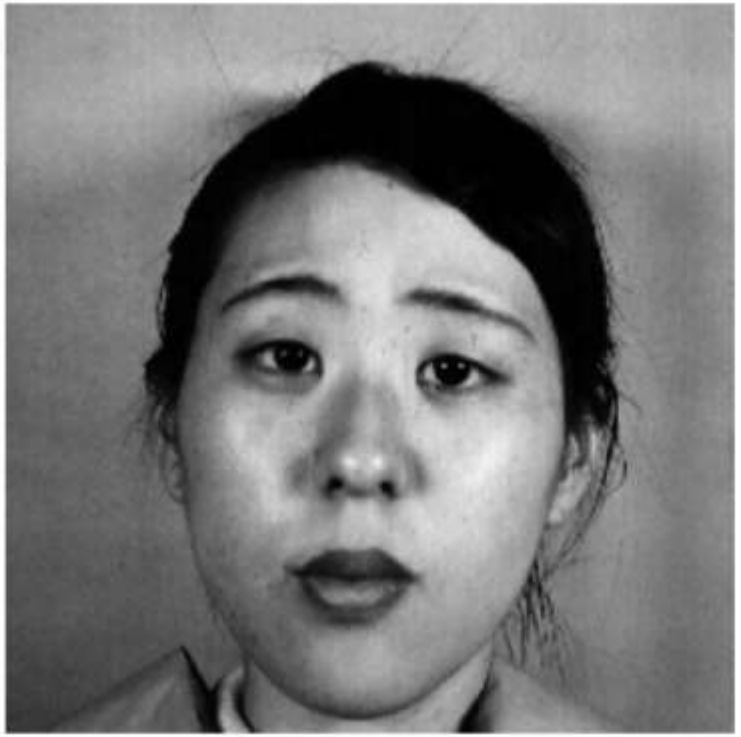


Figure 4. The visualization of a typical result of our method on the SJA_c dataset and its corresponding image.

Thanks