Label Distribution Learning

Yiming Wang
2022/05/11

Introduction

- Which label can describe the instance?

Cat / Dog

Single Label Learning (SLL)

Color?
Type?

Multi Label Learning (MLL)

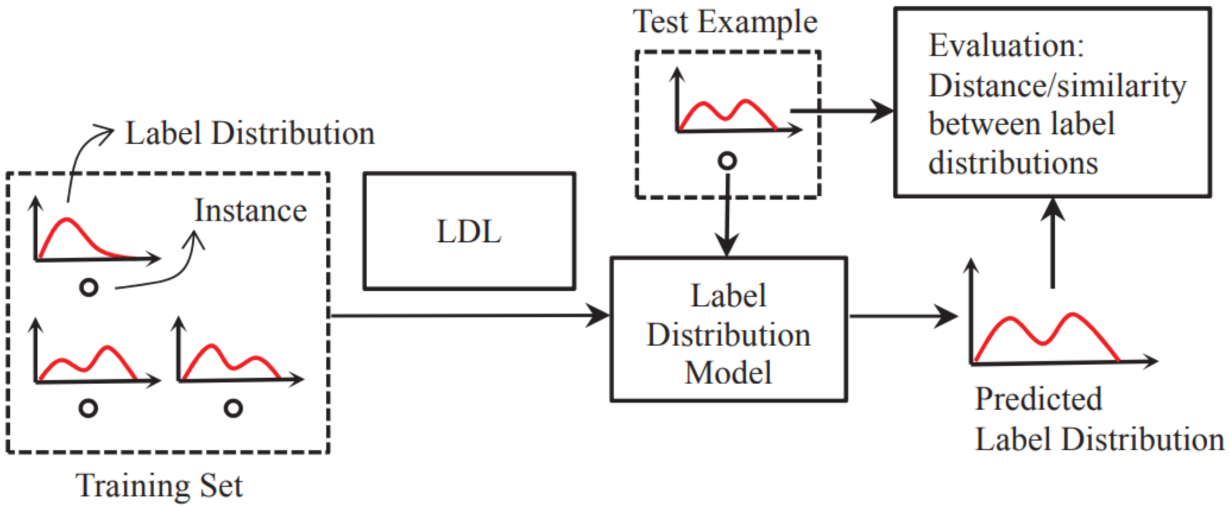
How much does each label describe the instance?

Figure 1. Four images with different emotion distributions from the involved datasets. Rather than a dominant emotion, images often evoke multiple emotions with different description degrees.

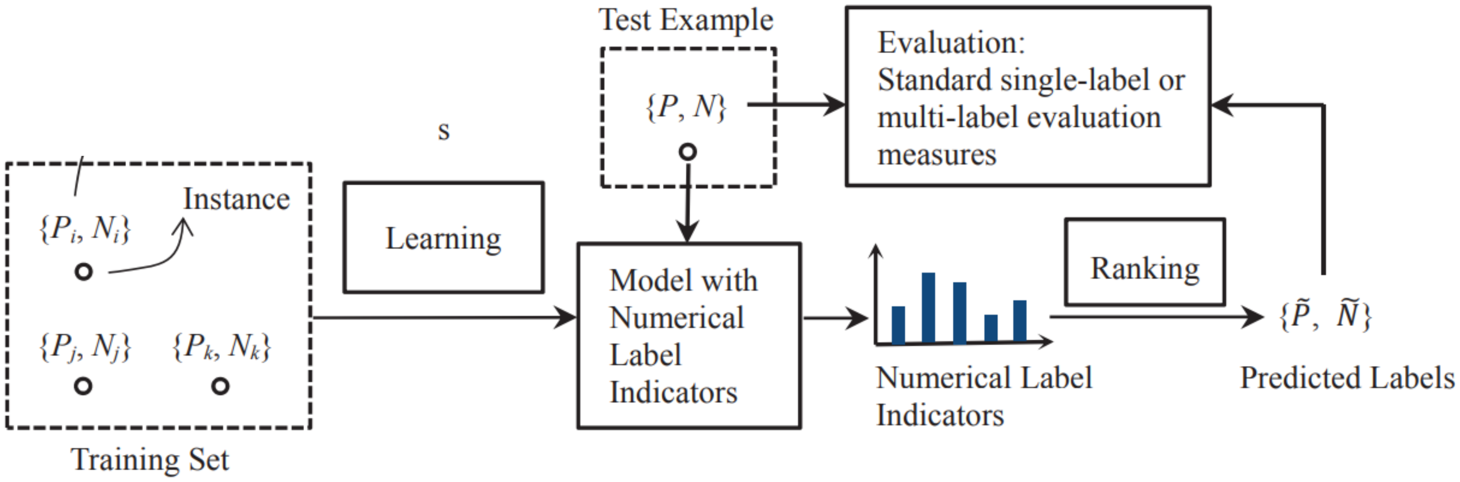
Label Distribution Learning (LDL)

Introduction

- Which label can describe the instance?[1]



(a) LDL



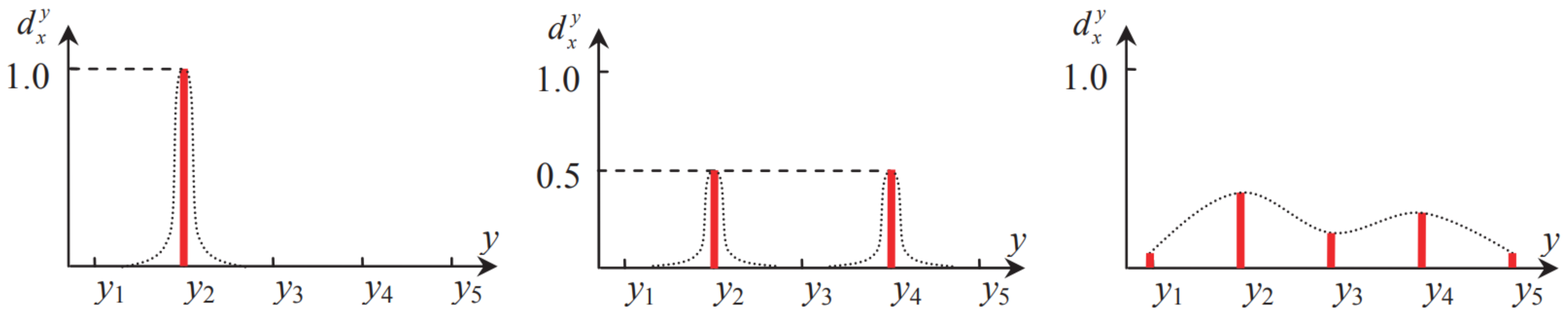
(b) Typical existing learning methods with numerical label indicators

[1] X. Geng et al. Label Distribution Learning. IEEE Transactions on Knowledge & Data Engineering. 2014.

Definition

- Suppose there is an **instance** $\boldsymbol{x}$, and we use d_x^y describe the **degree of** $\boldsymbol{x}$ **for the label** $\boldsymbol{y}$, so the **GT** of x can be $D_i = \{d_x^{y_1}, d_x^{y_2}, \ldots, d_x^{y_c}\}$, where c is the number of possible labels.
- If the following conditions are met, it is a Label Distribution Learning.

$$d_x^y \in [0, 1] \qquad\qquad \sum_y d_x^y = 1$$



(a) Single-label annot. (b) Multi-label annot. (c) General case

Formulation of LDL (1/2)

- d_x^y can also be represented by the form of conditional probability as $d_x^y = P(y|x)$. Then, the problem of LDL can be formulated as follows.

Let $\mathcal{X} = \mathbb{R}^q$ denote the input space and $\mathcal{Y} = \{y_1, y_2, \cdots, y_c\}$ denote the complete set of labels. Given a training set $S = \{(\boldsymbol{x}_1, D_1), (\boldsymbol{x}_2, D_2), \cdots, (\boldsymbol{x}_n, D_n)\}$, the goal of ldl is to learn a conditional probability mass function $p(y|\boldsymbol{x})$ from S, where $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$.

- Suppose $P(y|x)$ is a parametric model $P(y|x{:}\,\theta)$ where θ is the parameter vector. The goal of LDL is to find the θ that can generate a distribution which is the most similar one to the GT distribution.

Formulation of LDL (2/2)

- If KL loss is set as the loss function, the best parameter $\boldsymbol{\theta}^*$ is determined by:

$$\begin{aligned}\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_i \sum_j \left(d_{\boldsymbol{x}_i}^{y_j} \ln \frac{d_{\boldsymbol{x}_i}^{y_j}}{p(y_j|\boldsymbol{x}_i;\boldsymbol{\theta})} \right) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_i \sum_j d_{\boldsymbol{x}_i}^{y_j} \ln p(y_j|\boldsymbol{x}_i;\boldsymbol{\theta}).\end{aligned}$$

Since

Kullback-Leibler ↓	$Dis_4(D, \widehat{D}) = \sum_{j=1}^{c} d_j \ln \frac{d_j}{\hat{d}_j}$

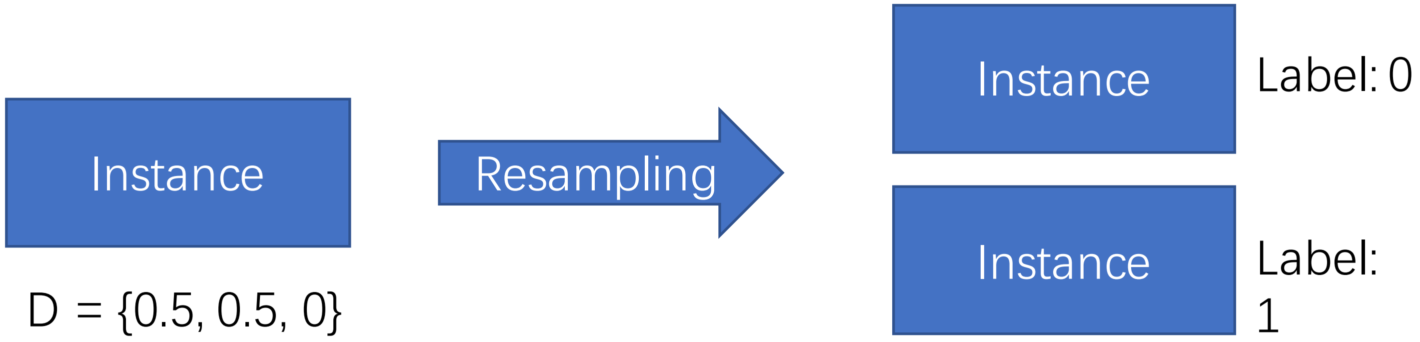
LDL Algorithms

- There are commonly three ways to solve LDL problems.
  - Problem transformation (PT)
    - Transfer LDL problems to existing learning paradigms. (SLL, MLL, etc.)
  - Algorithm adaptation (AA)
    - Extend existing algorithms to solve label distribution problems.
  - Specialized algorithms (SA)
    - Design specific algorithms for label distribution problems.

Problem Transformation

- **Change the training examples into weighted single-label examples.**

- For example, LDL -> SLL



- Each training example (x_i, D_i) is transformed into c single-label examples, where c is the number of classes, so there will be $i * c$ new samples as the training set.

- Then SSL algorithms (Bayes/SVM) can be used.

Algorithm Adaptation

- **Some traditional methods can be extended to solve LDL problems.**

- AA-kNN
  - ➢Choose the *k* nearest neighbors, then the distribution is calculated by counting the number of instances with each label.

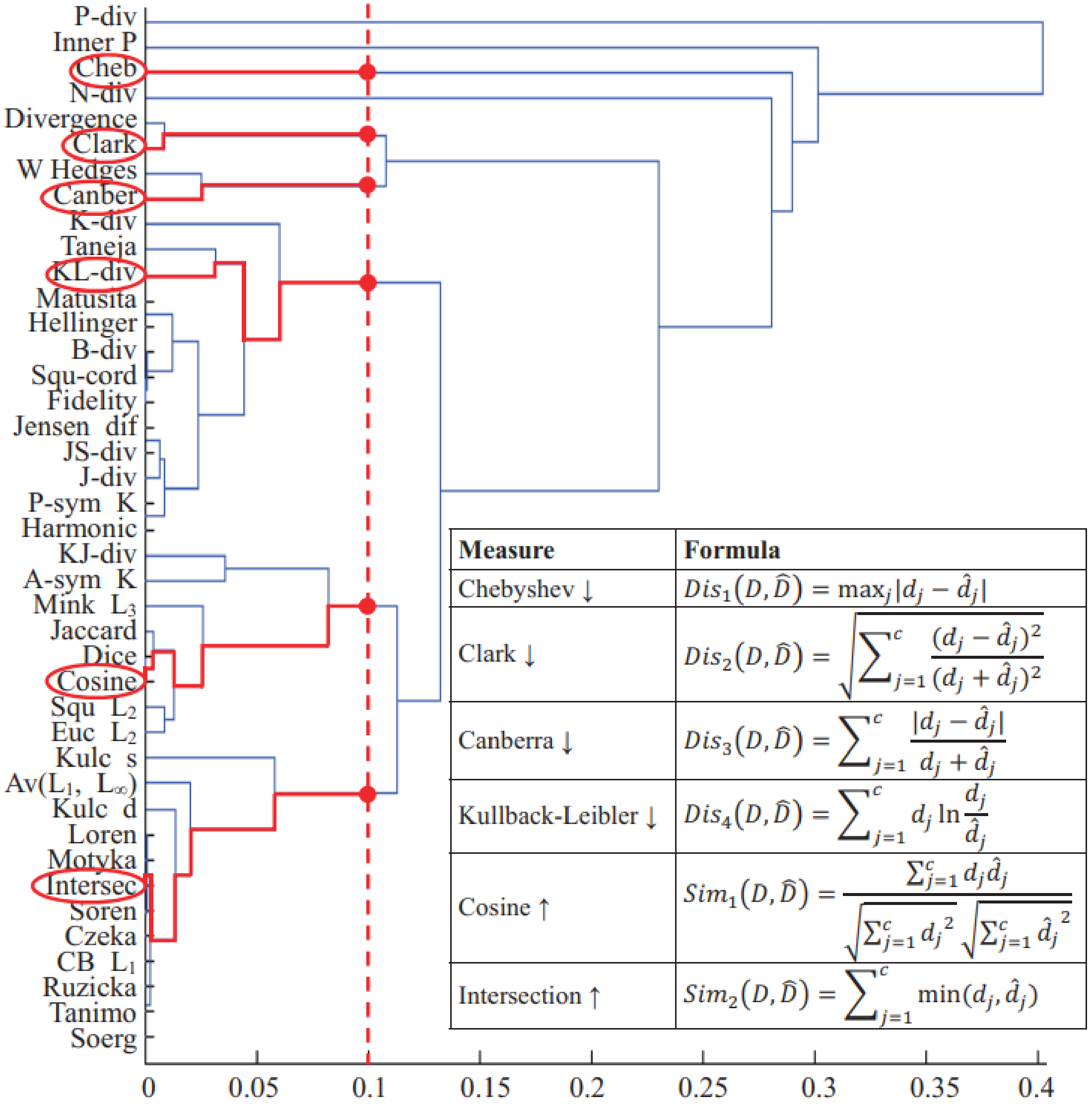
$$p(y_j|\boldsymbol{x}) = \frac{1}{k} \sum_{i \in N_k(\boldsymbol{x})} d_{\boldsymbol{x}_i}^{y_j}, (j = 1, 2, \ldots, c),$$

- **AA-B**ack**p**ropagation (BP)
  - ➢The output is not 0 or 1 but a probability.

Evaluation Measures

- To measure the output of LDL, people usually use the **similarity** or **distance** between the output and the GT.

Measure	Formula
Chebyshev ↓	$Dis_1(D,\hat{D}) = \max_j \lvert d_j - \hat{d}_j \rvert$
Clark ↓	$Dis_2(D,\hat{D}) = \sqrt{\sum_{j=1}^{c} \frac{(d_j - \hat{d}_j)^2}{(d_j + \hat{d}_j)^2}}$
Canberra ↓	$Dis_3(D,\hat{D}) = \sum_{j=1}^{c} \frac{\lvert d_j - \hat{d}_j \rvert}{d_j + \hat{d}_j}$
Kullback-Leibler ↓	$Dis_4(D,\hat{D}) = \sum_{j=1}^{c} d_j \ln \frac{d_j}{\hat{d}_j}$
Cosine ↑	$Sim_1(D,\hat{D}) = \frac{\sum_{j=1}^{c} d_j \hat{d}_j}{\sqrt{\sum_{j=1}^{c} {d_j}^2}\sqrt{\sum_{j=1}^{c} {\hat{d}_j}^2}}$
Intersection ↑	$Sim_2(D,\hat{D}) = \sum_{j=1}^{c} \min(d_j, \hat{d}_j)$



Datasets

- There are in total 16 datasets used in the experiments including **an artificial toy dataset** and **15 real-world datasets**.

TABLE 1
Statistics of the 16 Datasets Used in the Experiments

No.	Dataset	# Examples (n)	# Features (q)	# Labels (c)
1	Artificial	500 (train) 40,401 (test)	3	3
2	Yeast-alpha	2,465	24	18
3	Yeast-cdc	2,465	24	15
4	Yeast-elu	2,465	24	14
5	Yeast-diau	2,465	24	7
6	Yeast-heat	2,465	24	6
7	Yeast-spo	2,465	24	6
8	Yeast-cold	2,465	24	4
9	Yeast-dtt	2,465	24	4
10	Yeast-spo5	2,465	24	3
11	Yeast-spoem	2,465	24	2
12	Human Gene	30,542	36	68
13	Natural Scene	2,000	294	9
14	SJAFFE	213	243	6
15	SBU_3DFE	2,500	243	6
16	Movie	7,755	1,869	5

Datasets

- The first toy dataset is generated to show in a direct and visual way whether the LDL algorithms can learn the mapping from the instance to the label distribution.

No.	Dataset	# Examples (n)	# Features (q)	# Labels (c)
1	Artificial	500 (train) 40,401 (test)	3	3

$$
\begin{aligned}
t_i &= ax_i + bx_i^2 + cx_i^3 + d, i = 1, \ldots, 3, \\
\psi_1 &= (\boldsymbol{w}_1^{\mathrm{T}} \boldsymbol{t})^2, \\
\psi_2 &= (\boldsymbol{w}_2^{\mathrm{T}} \boldsymbol{t} + \lambda_1 \psi_1)^2, \\
\psi_3 &= (\boldsymbol{w}_3^{\mathrm{T}} \boldsymbol{t} + \lambda_2 \psi_2)^2, \\
d_{\boldsymbol{x}}^{y_i} &= \frac{\psi_i}{\psi_1 + \psi_2 + \psi_3}, i = 1, \ldots, 3,
\end{aligned}
$$

a = 1, b = 0.5, c = 0.2, each component x is uniformly sampled within the range [-1 1]

$\boldsymbol{w}_1 = [4, 2, 1]^{\mathrm{T}}$, $\boldsymbol{w}_2 = [1, 2, 4]^{\mathrm{T}}$ $\boldsymbol{w}_3 = [1, 4, 2]^{\mathrm{T}}$

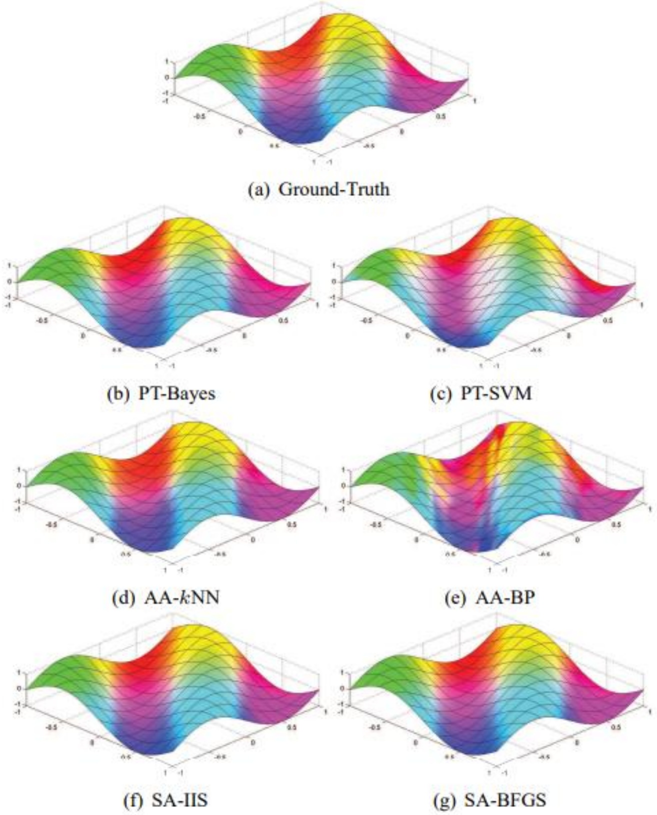


Fig. 4. Comparison between the ground-truth and predicted label distributions (regarded as RGB colors) on the artificial test manifold.

Experiment

- There are 6 measurements to show the performance of LDL algorithms. Here is the KL Loss case.

TABLE 7
Experimental Results (mean±std(rank)) on the Real-world Datasets Measured by Kullback-Leibler Divergence ↓

Dataset	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS
Yeast-alpha	0.719±0.080(6)	0.009±0.002(4)	0.0066±0.001(2)	0.081±0.011(5)	0.0067±0.001(3)	**0.006±0.001(1)**
Yeast-cdc	0.603±0.073(6)	0.010±0.002(4)	0.0083±0.001(3)	0.060±0.007(5)	0.0082±0.001(2)	**0.007±0.001(1)**
Yeast-elu	0.556±0.071(6)	0.008±0.001(4)	0.0074±0.0004(3)	0.051±0.009(5)	0.0073±0.0005(2)	**0.006±0.0004(1)**
Yeast-diau	0.306±0.036(6)	0.019±0.002(4)	0.015±0.001(3)	0.024±0.004(5)	0.014±0.001(2)	**0.013±0.001(1)**
Yeast-heat	0.255±0.040(6)	0.0148±0.001(4)	0.0145±0.001(3)	0.021±0.004(5)	0.0133±0.0004(2)	**0.0126±0.0005(1)**
Yeast-spo	0.281±0.031(6)	0.0304±0.005(4)	0.0302±0.002(3)	0.034±0.006(5)	0.0254±0.003(2)	**0.0246±0.003(1)**
Yeast-cold	0.208±0.031(6)	0.0147±0.001(4)	0.014±0.001(3)	0.0149±0.002(5)	0.013±0.001(2)	**0.012±0.001(1)**
Yeast-dtt	0.206±0.029(6)	0.0073±0.001(4)	0.0072±0.001(3)	0.009±0.001(5)	0.0070±0.001(2)	**0.006±0.001(1)**
Yeast-spo5	0.214±0.025(6)	0.03010±0.003(3)	0.033±0.003(5)	0.031±0.003(4)	0.03007±0.003(2)	**0.029±0.003(1)**
Yeast-spoem	0.190±0.038(6)	0.0280±0.004(4)	0.0285±0.003(5)	0.026±0.003(3)	0.025±0.003(2)	**0.024±0.003(1)**
Human Gene	1.887±0.766(6)	0.240±0.019(3)	0.301±0.026(4)	0.500±0.068(5)	0.238±0.019(2)	**0.236±0.019(1)**
Natural Scene	3.065±0.487(6)	1.447±0.243(4)	2.767±0.137(5)	0.875±0.029(3)	0.870±0.026(2)	**0.854±0.062(1)**
s-JAFFE	0.074±0.014(4)	0.086±0.016(5)	0.071±0.023(3)	0.113±0.030(6)	0.070±0.012(2)	**0.064±0.016(1)**
s-BU_3DFE	0.079±0.004(4)	0.089±0.007(6)	0.065±0.002(2)	0.085±0.009(5)	0.068±0.004(3)	**0.049±0.002(1)**
Movie	0.953±0.352(6)	0.268±0.079(5)	0.201±0.011(4)	0.179±0.03(3)	**0.137±0.013(1)**	0.140±0.020(2)
Avg. Rank	5.73	4.13	3.40	4.60	2.07	1.07

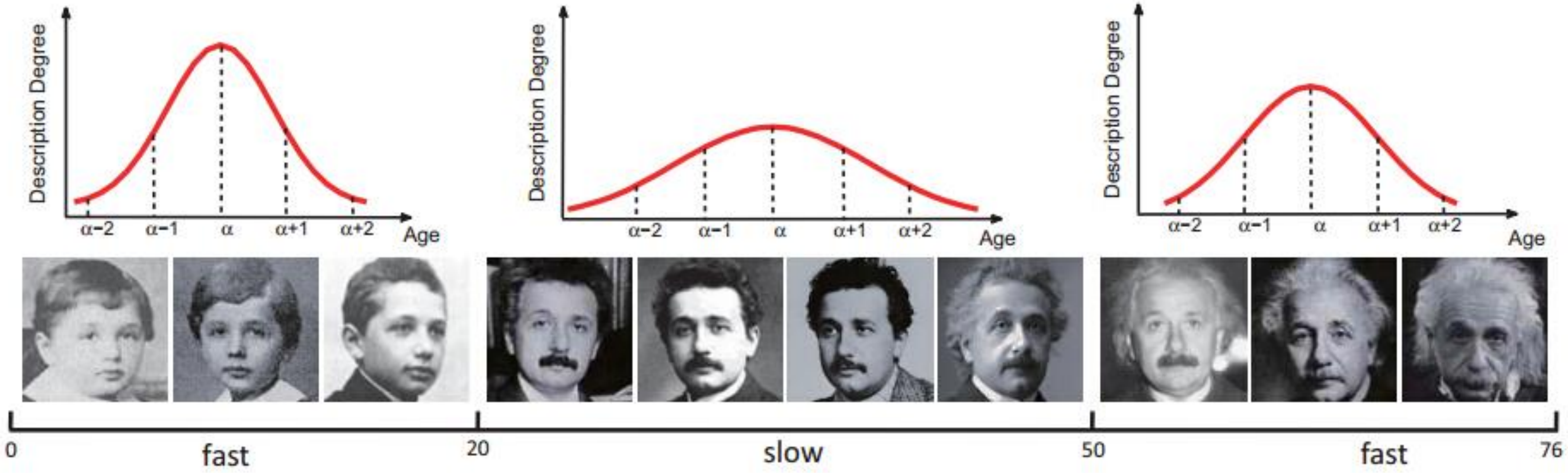
In which cases can we try LDL?

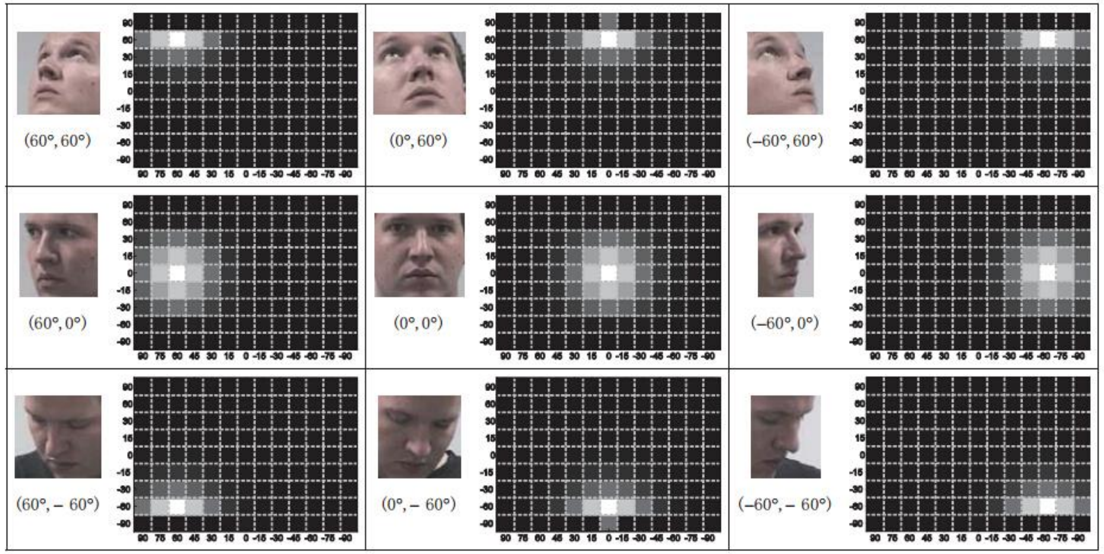
- There are intrinsic relationships between labels.
  - ➢Emotion recognition
- The labels are Subjectivity and Ambiguity (No Exact GT)
  - ➢Rating estimation[2]
  - ➢Aesthetics estimation
- Other estimation studies.
  - ➢Number counting[3]
  - ➢Even imagability, Bouba/Kiki?

[2] X. Geng et al. Pre-release Prediction of Crowd Opinion on Movies by Label Distribution Learning. IJCAI 2015.
[3] X. Wu. et al. Joint Acne Image Grading and Counting via Label Distribution Learning. ICCV 2019.

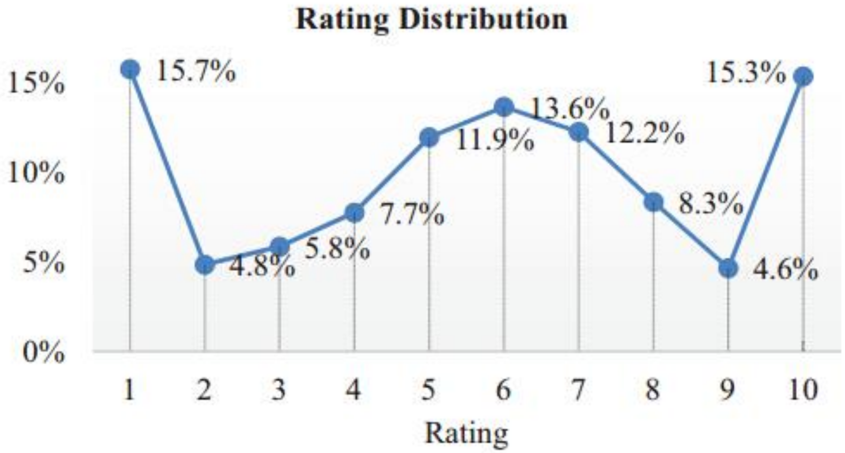
Applications of LDL[4]

- Facial Age Estimation
- Head Pose Estimation
- Pre-release Prediction of Movies
- ……





Title	Twilight
Average Rating	5.2/10
Budget	$ 37 Million
Gross	$ 191 Million



[4]http://palm.seu.edu.cn/xgeng/LDL/index.htm

Emotion Recognition via LDL[5]

- Visual sentiment analysis is ambiguous since an image usually evokes multiple emotions (Ambiguity) and its annotation varies from person to person (Subjectivity) --> LDL
- **Convert single label to distribution using two constraints.**
  - **Implication**: an emotion evokes related emotions.
  - **Exclusion**: positive/negative emotions do not evoke negative/positive emotions.

$$f(x, \mu, \sigma_{\text{conf}}) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_{\text{conf}}} exp\left(-\frac{|i-\mu|^2}{2\sigma_{\text{conf}}^2}\right), i \in Y_\mu \\ 0, i \notin Y_\mu \end{cases}$$

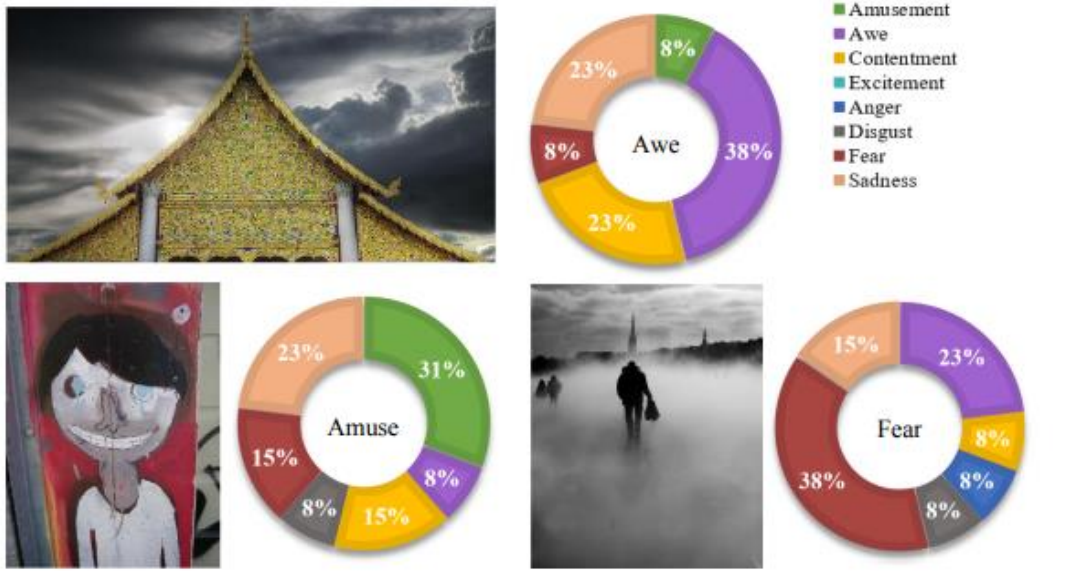


Figure 1: Images from the Flickr_LDL dataset are annotated by 11 users on 8 emotions. The pie chart on the right of each image demonstrates the label ambiguity. The dominant sentiment of each image is also shown.

[5]Yang J. et al. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. IJCAI 2017.

Emotion Recognition Framework

VGG-16

$$L = (1 - \lambda) L_{cls}(x, y) + \lambda L_{sdl}(x, \boldsymbol{l}),$$

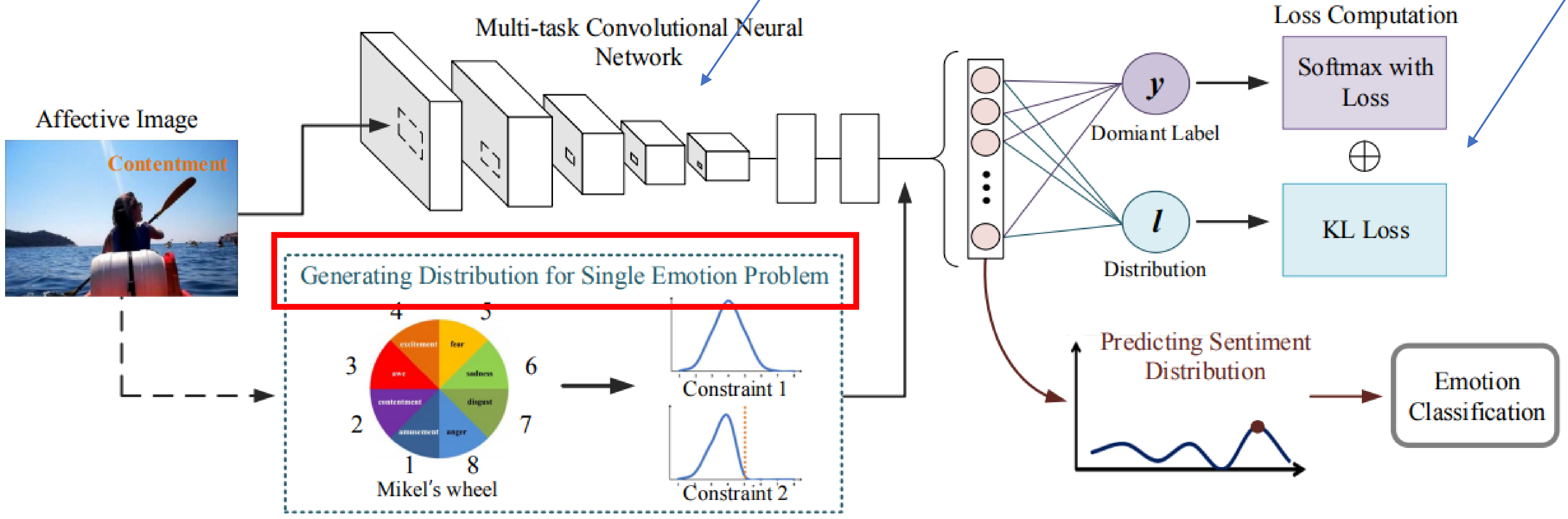


Figure 2: The illustration of our method. Given the affective images with distribution, our framework simultaneously optimize the classification loss and distribution loss. In details, the softmax loss is employed as the classification constraints, while the KL loss is added for distribution learning. For the single emotion dataset, we also propose to transform single label into label distribution according to two weak prior knowledge.

Emotion Recognition Experiment

Table 1: Experimental Results on three distribution datasets, *i.e.* Emotion6 (E), Flickr_LDL (F), and Twitter_LDL (T), are shown as mean(rank). Since each measure reflects a certain aspect of an algorithm, "Avg Rank" is used to indicate the overall performance of distribution prediction. "Acc" indicates the classification result of the single dominant emotional category.

	Criterion	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS	SA-CPNN	BCPNN	ACPNN	CNNR	DLDL	Ours
E	Cheb $\downarrow$	0.35(10)	0.39(12)	0.29(6)	0.30(7)	0.32(9)	0.38(11)	0.30(7)	0.28(5)	0.27(4)	0.26(3)	0.25(2)	**0.24(1)**
	Clark $\downarrow$	1.94(11)	1.82(10)	1.63(3)	1.69(9)	1.67(7)	1.96(12)	1.68(8)	1.66(6)	1.66(5)	**1.61(1)**	1.64(4)	1.62(2)
	Canber $\downarrow$	4.59(11)	4.31(10)	3.60(3)	3.79(8)	3.83(9)	4.68(12)	3.78(7)	3.73(6)	3.68(5)	**3.46(1)**	3.63(4)	3.58(2)
	KLdiv $\downarrow$	2.32(12)	1.07(10)	0.85(9)	0.63(7)	0.61(6)	1.16(11)	0.56(5)	0.52(4)	0.50(3)	0.67(8)	0.43(2)	**0.42(1)**
	Cosine $\uparrow$	0.69(8)	0.48(12)	0.75(4)	0.68(9)	0.69(7)	0.63(11)	0.66(10)	0.75(5)	0.76(3)	0.74(6)	0.79(2)	**0.80(1)**
	Intersec $\uparrow$	0.56(10)	0.42(12)	0.62(4)	0.59(9)	0.61(6)	0.52(11)	0.60(8)	0.62(5)	0.63(3)	0.60(7)	0.65(2)	**0.65(1)**
	Avg Rank	10.3(10)	11.0(11)	4.83(5)	8.17(9)	7.33(7)	11.3(12)	7.50(8)	5.17(6)	3.83(3)	4.33(4)	2.67(2)	**1.33(1)**
	Acc.(%)	39.2(10)	36.7(11)	44.1(6)	39.5(9)	41.1(8)	34.6(12)	42.2(7)	45.4(4)	46.9(2)	45.2(4)	46.1(3)	**52.4(1)**
F	Cheb $\downarrow$	0.44(11)	0.55(12)	0.28(6)	0.36(9)	0.31(8)	0.37(10)	0.30(7)	0.28(5)	0.25(4)	0.25(3)	0.25(2)	**0.24(1)**
	Clark $\downarrow$	2.51(12)	2.45(11)	**1.62(1)**	2.33(8)	2.33(9)	2.44(10)	2.31(7)	2.21(4)	2.19(3)	2.29(6)	2.22(5)	2.19(2)
	Canber $\downarrow$	6.76(12)	6.61(11)	**3.30(1)**	5.98(8)	6.00(9)	6.44(10)	5.91(7)	5.63(5)	5.57(3)	5.82(6)	5.59(4)	5.55(2)
	KLdiv $\downarrow$	1.88(11)	1.69(10)	3.28(12)	0.82(8)	0.66(5)	1.06(9)	0.71(7)	0.62(4)	0.61(3)	0.70(6)	0.54(2)	**0.53(1)**
	Cosine $\uparrow$	0.63(11)	0.32(12)	0.79(5)	0.72(8)	0.78(6)	0.70(10)	0.70(9)	0.80(4)	0.81(3)	0.72(7)	0.81(2)	**0.82(1)**
	Intersec $\uparrow$	0.49(11)	0.29(12)	0.64(3)	0.53(9)	0.60(7)	0.56(8)	0.60(6)	0.62(5)	0.63(4)	0.62(10)	0.64(2)	**0.65(1)**
	Avg Rank	11.3(11)	11.3(11)	4.67(5)	8.33(9)	7.33(8)	9.50(10)	7.17(7)	4.50(4)	3.33(3)	6.33(6)	2.83(2)	**1.33(1)**
	Acc.(%)	46.9(11)	37.3(12)	61.4(2)	52.0(9)	57.9(7)	50.1(10)	57.7(8)	59.7(6)	60.0(5)	60.7(4)	60.9(3)	**64.2(1)**
T	Cheb $\downarrow$	0.53(11)	0.63(12)	0.28(5)	0.31(8)	0.28(6)	0.37(10)	0.36(9)	0.31(7)	0.27(3)	0.28(4)	0.26(2)	**0.25(1)**
	Clark $\downarrow$	2.39(6)	2.56(12)	**1.65(1)**	2.40(8)	2.42(10)	2.51(11)	2.41(9)	2.38(4)	2.40(7)	2.37(3)	2.38(5)	2.36(2)
	Canber $\downarrow$	6.17(6)	7.05(12)	**3.30(1)**	6.26(9)	6.32(10)	6.70(11)	6.22(8)	6.15(4)	6.22(7)	6.11(3)	6.17(5)	6.05(2)
	KLdiv $\downarrow$	1.31(10)	1.65(11)	3.89(12)	0.68(7)	0.64(5)	1.19(9)	0.85(8)	0.61(4)	0.58(3)	0.67(6)	0.54(2)	**0.53(1)**
	Cosine $\uparrow$	0.53(11)	0.25(12)	0.82(5)	0.81(8)	0.82(6)	0.71(10)	0.75(9)	0.83(4)	0.84(2)	0.82(7)	0.83(3)	**0.85(1)**
	Intersec $\uparrow$	0.40(11)	0.21(12)	0.66(2)	0.59(7)	0.63(5)	0.57(9)	0.56(10)	0.60(6)	0.64(4)	0.58(8)	0.65(3)	**0.68(1)**
	Avg Rank	9.17(10)	11.8(12)	4.33(3)	7.83(8)	7.00(7)	10.0(11)	8.83(9)	4.83(5)	4.33(3)	5.17(6)	3.33(2)	**1.33(1)**
	Acc.(%)	45.1(11)	40.4(12)	72.6(5)	72.4(7)	70.3(8)	57.0(10)	70.0(9)	73.0(4)	74.2(2)	73.6(3)	72.6(5)	**76.3(1)**

Emotion Recognition Ablation Study

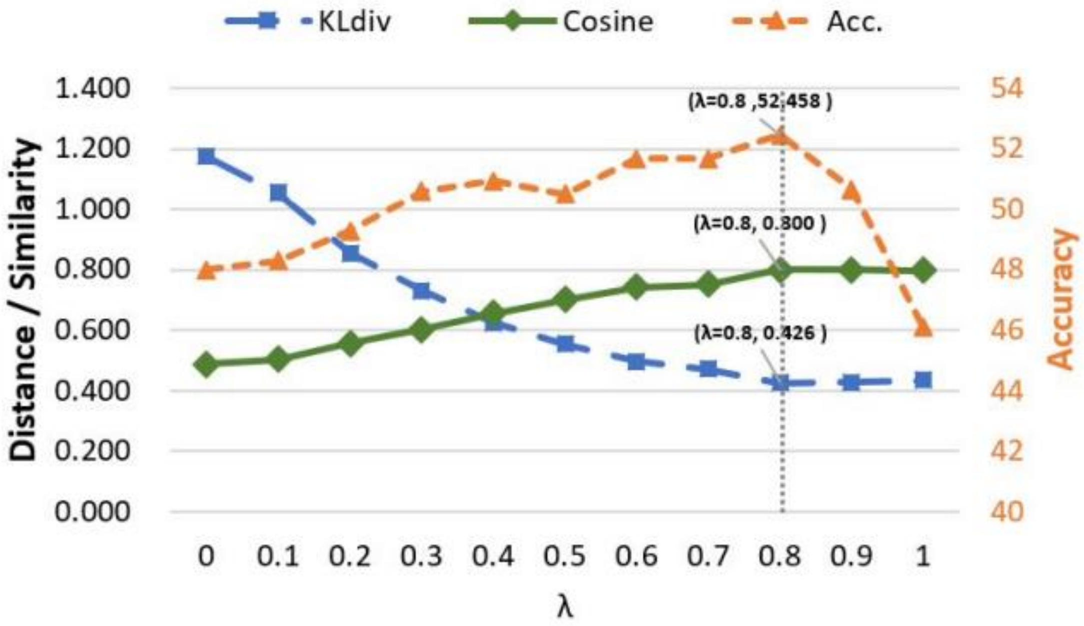


Figure 3: Effect of λ on the Emotion6 dataset, which indicates the weight of the distribution term in the optimization objective function. Note that $\lambda = 0$ represents that only softmax loss for classification is employed.

Table 2: Classification performance on the FI dataset.

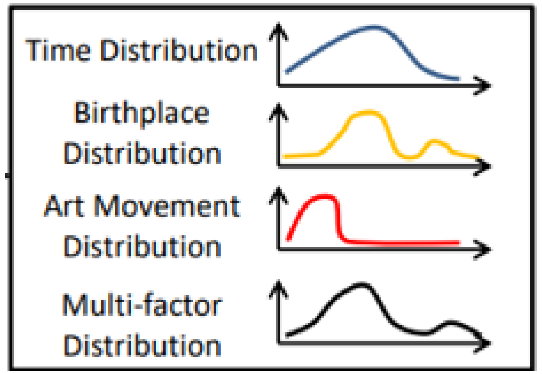
	Methods	Accuracy
Baseline	Zhao's	46.13%
	DeepSentiBank	51.54%
	PCNN (VGGNet)	55.24%
CNNs	AlexNet	41.28%
	VGGNet	46.22%
	ResNet	49.76%
	Fine-tuned AlexNet	58.13%
	Fine-tuned VGGNet	63.75%
	Fine-tuned ResNet	64.67%
Ours	ours (AlexNet)	**60.63%**
	ours (VGGNet)	**66.21%**
	ours (ResNet)	**66.79%**
	ours (Ensemble)	**67.48%**

Table 3: Comparison of different methods for emotion classification on the FI dataset.

Methods	Accuracy
VGG (softmax, $\lambda = 0$)	63.75%
VGG + Constraint1 ($\lambda = 0.6$)	66.00%
VGG + Constraint2 ($\lambda = 0.6$)	65.18%
VGG + Constraint1 ($\lambda = 0.8$)	66.21%
VGG + Constraint2 ($\lambda = 0.8$)	65.27%
VGG + Constraint1 (KL-div, $\lambda = 1$)	64.95%
VGG + Constraint2 (KL-div, $\lambda = 1$)	64.28%
VGG + LS ($\lambda = 0.8$)	64.15%

Art Style Classification via LDL[6]

- There are some intrinsic relationships between different art styles. For example, one style may inherit from anther style. Therefore, LDL can also be applied here.
- Multi factor distribution
  - Time distribution
  - Birthplace distribution
  - Art movement distribution



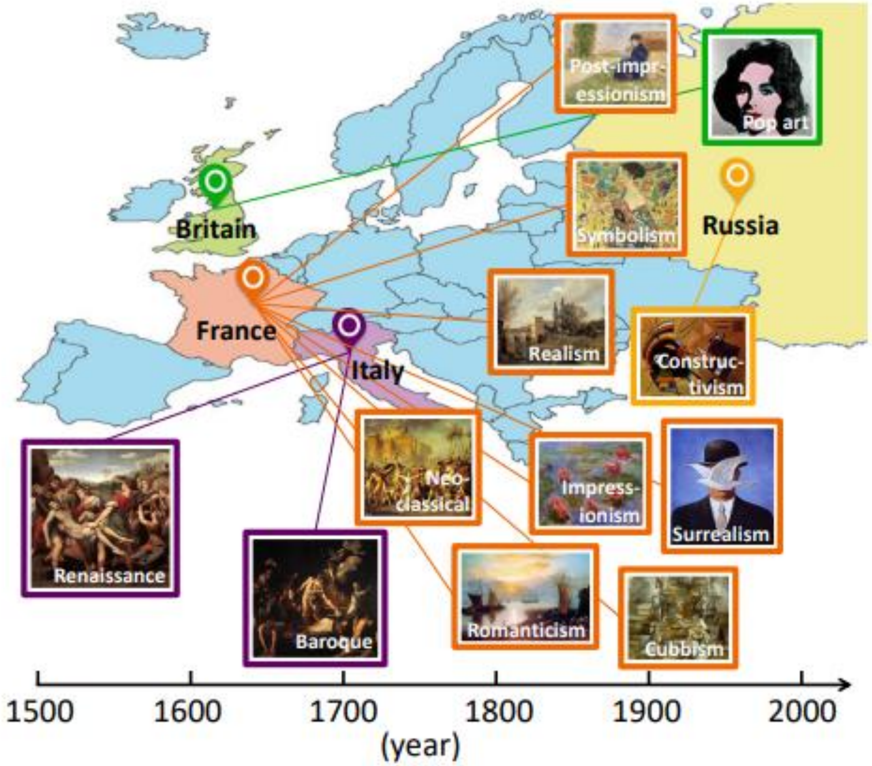
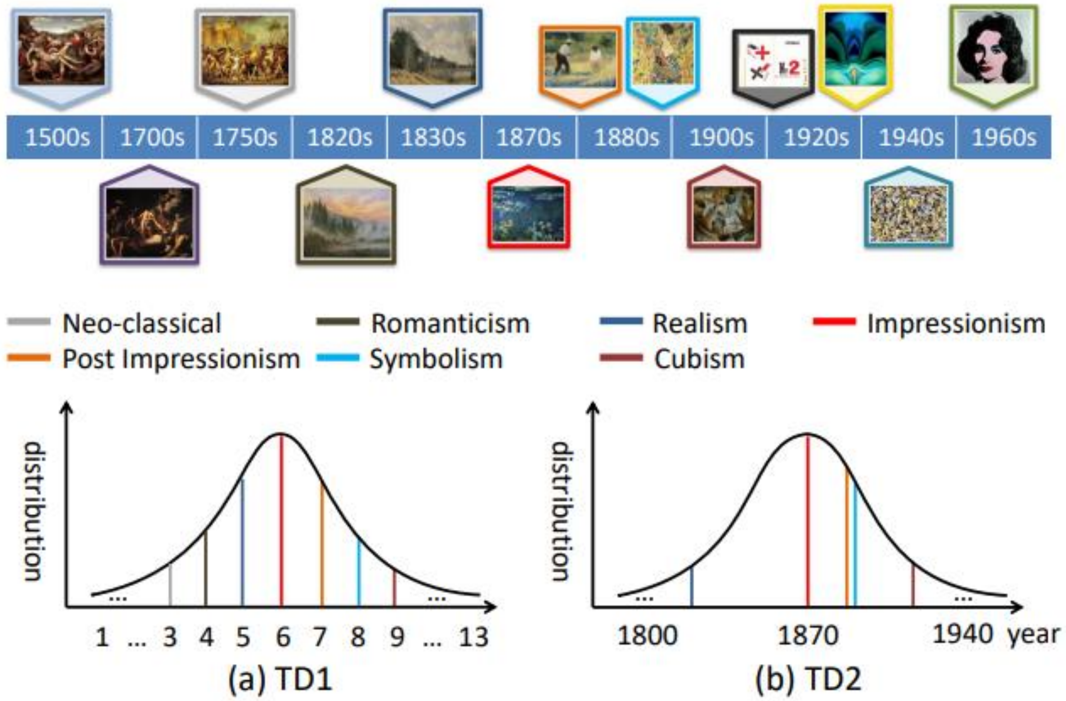


Figure 1: The origin time and birthplace of painting styles in the Painting91 dataset. There are some paintings from different styles and we arrange them according to the chronological order, as indicated by the horizontal axis.

[6]Yang J. et al. Historical Context-based Style Classification of Painting Images via Label Distribution Learning. ACM MM 2018.

Art Style Distribution



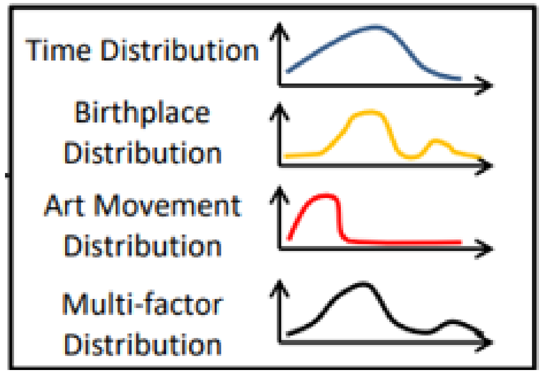
1.Time distribution

$$\boldsymbol{b}_i = \begin{cases} 1, & i = y \\ \frac{\beta}{n_b}, & B_i = B_y, i \neq y \\ 0, & otherwise \end{cases},$$

2.Birthplace distribution

$$\boldsymbol{a}_i = \begin{cases} 1, & i = y \\ \frac{\alpha}{n_a}, & A_i = A_y, i \neq y \\ 0, & otherwise \end{cases},$$

3.Art movement distribution



Multi factor distribution:

$$\boldsymbol{l} = \eta \times \boldsymbol{t1} + (1 - \eta) \times \boldsymbol{t2} + \boldsymbol{b} + \boldsymbol{a},$$

Art Style Framework

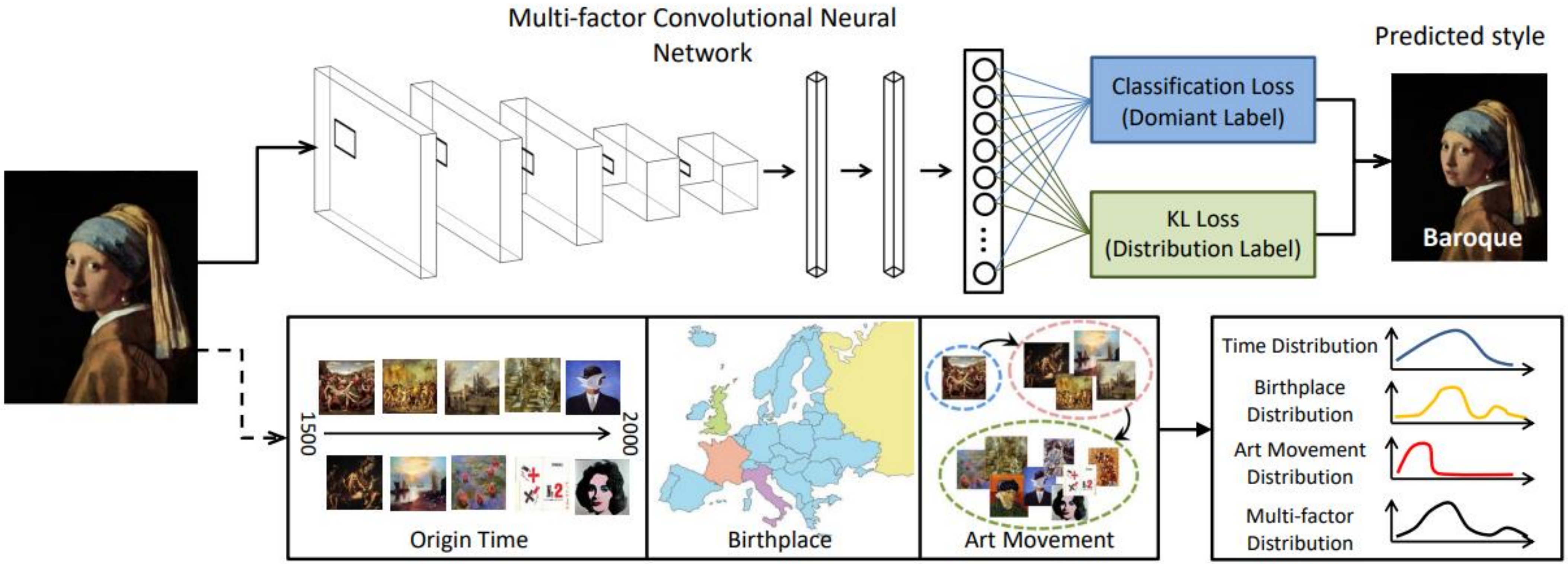


Figure 2: The illustration of the proposed method. Taking into account the three factors in the historical context that describe the relationship between styles (Origin Time, Birthplace, Art Movement), the framework simultaneously optimizes the classification loss and distribution loss. The softmax loss is employed as the classification loss, while the style distribution loss (KL loss) is used as an auxiliary task to assist visual feature learning towards better generalization ability.

Art Style Experiment

Table 1: Ablation experiments on the Painting91, OilPainting, and Pandora datasets. The first line denotes baseline using the single label. And we consider four additional properties of historical context with different label distributions. Note that TD1, TD2, BP, and AM represent two time distribution strategies, Birthplace, and Art movement, respectively.

Base	TD1	TD2	BP	AM	Painting91	OilPainting	Pandora
√					72.89%	64.24%	70.52%
	√				76.29%	69.58%	71.09%
		√			75.93%	68.88%	71.12%
			√		76.66%	69.28%	72.21%
				√	76.38%	69.05%	71.95%
	√	√			77.11%	69.85%	71.20%
	√	√	√		77.39%	70.23%	72.87%
	√	√		√	77.21%	70.10%	72.53%
	√	√	√	√	**77.76%**	**70.59%**	**73.28%**

Table 2: Classification performance on the test set of Painting91 dataset, OilPainting dataset, and Pandora dataset. Note that some methods do not provide the source code, thus some datasets cannot be evaluated, denoted as '-'.

Method	Painting91	OilPainting	Pandora
VGGNet [44]	72.89%	64.24%	70.52%
Khan F. S. *et al.* [23]	62.20%	-	-
Condorovici *et al.* [6]	-	-	37.90%
Florea *et al.* [9]	-	-	54.70%
CMFFV [37]	67.43%	-	-
MSCNN1 [34]	69.67%	55.24%	70.32%
MSCNN2 [34]	70.96%	57.92%	69.75%
CNN F4 [33]	69.21%	58.47%	70.47%
Peng K. C. *et al.* [35]	71.05%	-	-
Gram [5]	71.86%	60.61%	-
Gram-Cov [5]	72.41%	60.72%	-
Gram dot Cos [5]	73.59%	63.33%	-
SCMFA [38]	73.16%	-	-
Anwer R. M. *et al.* [1]	74.80%	-	-
Ours	**77.76%**	**70.59%**	**73.28%**

Art Style Visualization

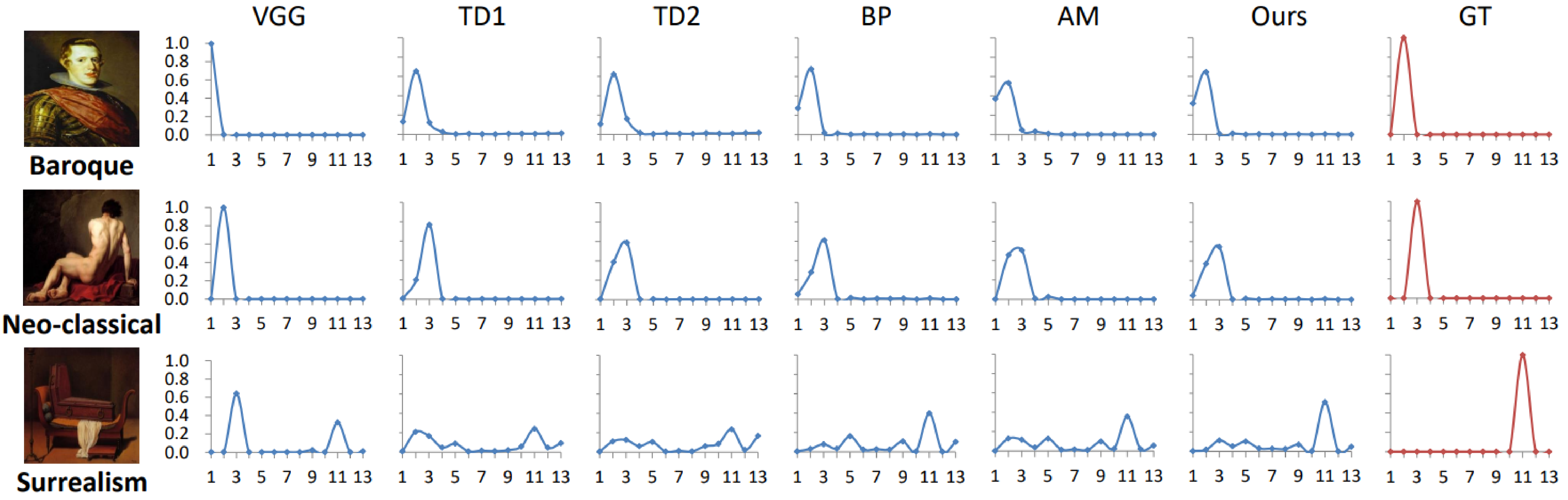


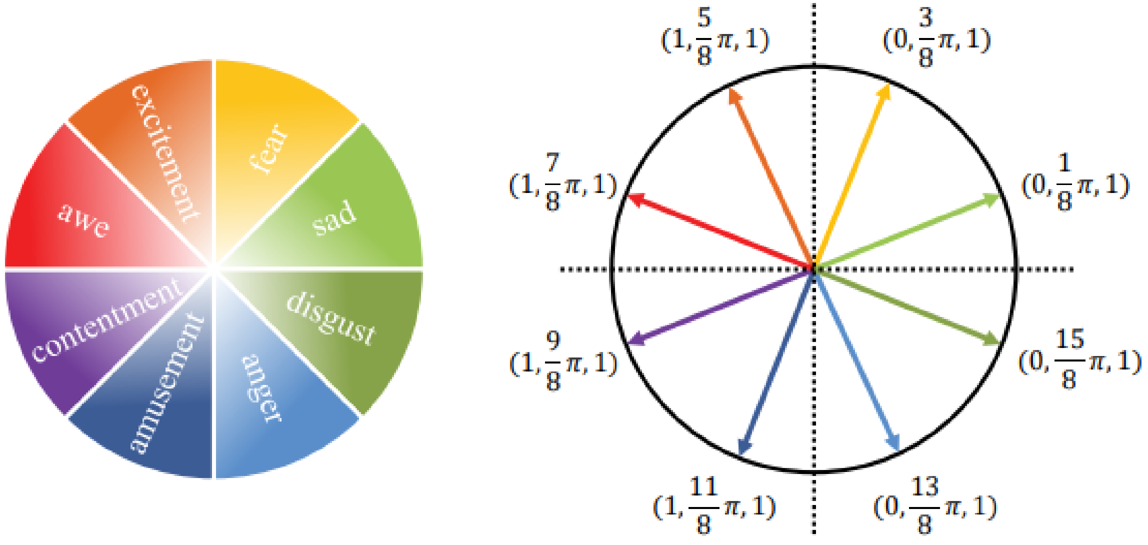
Figure 5: Examples from the Painting91 dataset with the predicted label distribution by VGGNet and our methods. For each subfigure, we introduce the style information at the bottom of the painting. On the right side of the painting, we list six predicted results using single label (VGG) and different label distribution methods (including two time strategies TD1 and TD2, the birthplace distribution (BP), the art movement distribution (AM), and multiple historical context factors (Ours)). The ground truth label (GT) is shown in the last column.

- Note that the GT is single labeled, but the prediction is a distribution.

New Constraints

- Some studies try to use the exist intrinsic relationships between classes.

- [7] uses the relationships between different emotions and give a new loss function.



(a) Mikel's Wheel (b) Emotion Circle

Figure 2. Mikel's Wheel from psychological model (a), and the proposed Emotion Circle (b) with eight basic emotion vectors evenly distributed in accordance with Mikel's Wheel.

$$\hat{\mathbf{e}}_i = (\hat{p}_i, \hat{\theta}_i, \hat{r}_i)$$

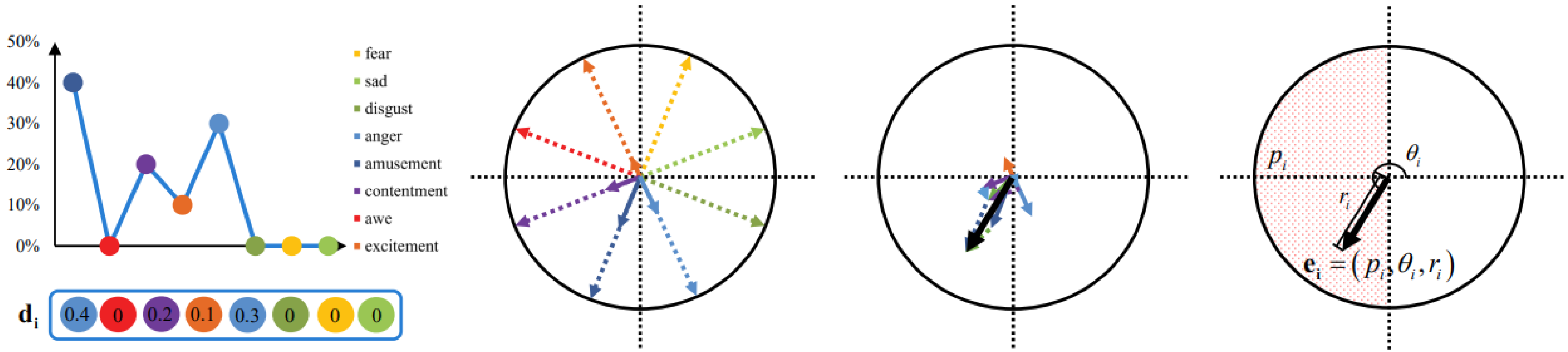
Emotion polarity: $\hat{p}_i$

Emotion type: $\hat{\theta}_i$

Emotion intensity: $\hat{r}_i$

[7]Yang J., et al. A Circular-Structured Representation for Visual Emotion Distribution Learning. CVPR 2021

Distribution to Emotion Vector



(a) Emotion distribution (b) Weight (c) Combination (d) Emotion vector

Figure 4. Mapping from the emotion distribution (a) to the compound emotion vector (d) on the Emotion Circle. We first weigh eight basic emotions with different description degrees (b) and then combine them to form a compound emotion vector through vector addition operations (c). The final emotion vector can be viewed as a specific circular-structured representation of a given emotion distribution.

Framework

$$\mathcal{L}_{PC} = \frac{1}{N}\sum_{i=1}^{N} r_i \left((p_i - \hat{p}_i)^2 + (\theta_i - \hat{\theta}_i)^2\right). \quad \mathcal{L} = (1-\mu)\mathcal{L}_{KL} + \mu\mathcal{L}_{PC},$$

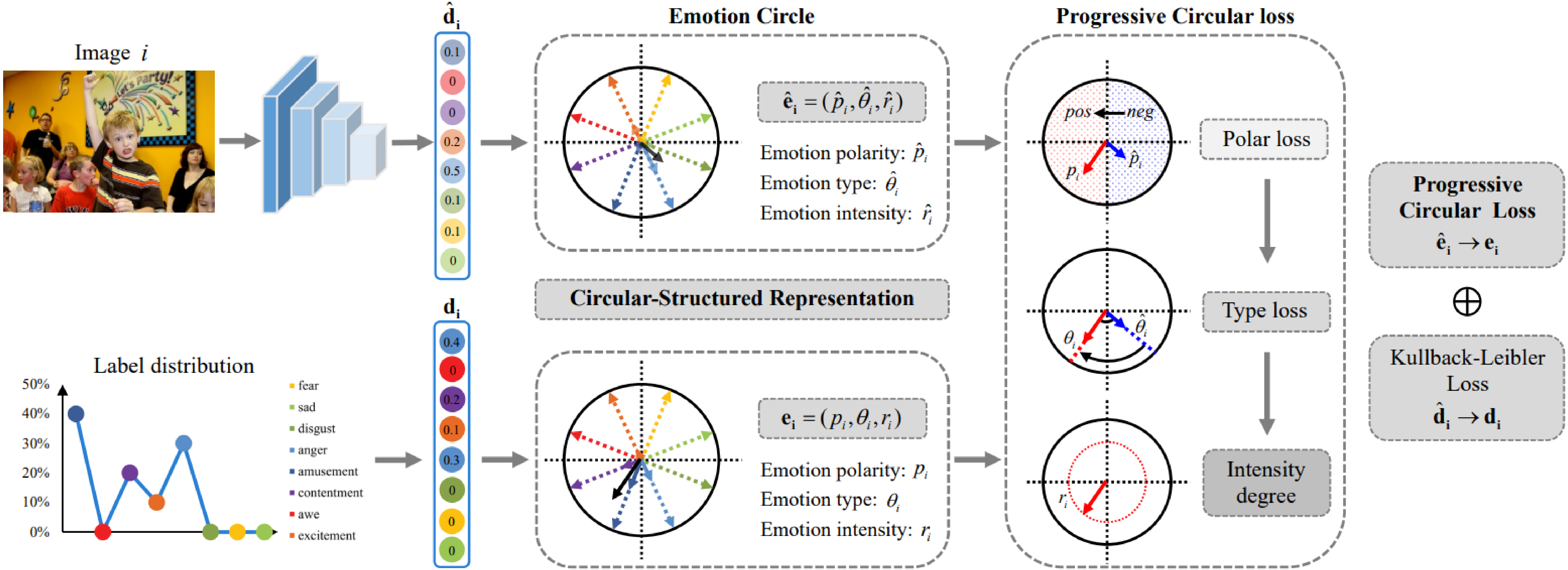


Figure 3. Framework of the proposed circular-structured representation. On the proposed Emotion Circle, both the predicted emotion distribution and the labeled one are represented with compound emotion vectors through a systematic approach. We then propose the Progressive Circular loss in a coarse-to-fine manner, which is further exploited to train the network together with Kullback-Leibler loss.

Experiment

Table 1. Comparison with the state-of-the-art methods on Flickr_LDL dataset.

	PT		AA		SA			CNN-based						
Measures	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS	SA-CPNN	CNNR	DLDL	ACPNN	JCDL	SSDL	E-GCN	Ours
Chebyshev ↓	0.44(13)	0.55(14)	0.28(8)	0.36(11)	0.31(10)	0.37(12)	0.30(9)	0.25(5)	0.25(5)	0.25(5)	0.24(4)	0.23(2)	0.23(2)	**0.21(1)**
Clark ↓	0.89(14)	0.87(13)	**0.57(1)**	0.82(8)	0.82(8)	0.86(12)	0.82(8)	0.84(11)	0.78(5)	0.77(2)	0.77(2)	0.78(5)	0.78(5)	0.77(2)
Canberra ↓	0.85(14)	0.83(13)	**0.41(1)**	0.75(10)	0.75(10)	0.82(12)	0.74(9)	0.73(8)	0.70(7)	0.70(5)	0.70(5)	0.69(3)	0.69(3)	0.68(2)
KL ↓	1.88(13)	1.69(12)	3.28(14)	0.82(10)	0.66(7)	1.06(11)	0.71(9)	0.70(8)	0.54(5)	0.62(6)	0.53(4)	0.46(3)	0.44(2)	**0.41(1)**
Cosine ↑	0.63(13)	0.32(14)	0.79(7)	0.72(9)	0.78(8)	0.70(11)	0.70(11)	0.72(9)	0.81(5)	0.80(6)	0.82(4)	0.85(3)	0.86(2)	**0.87(1)**
Intersection ↑	0.49(13)	0.29(14)	0.64(5)	0.53(12)	0.60(9)	0.56(11)	0.60(9)	0.62(7)	0.64(5)	0.62(7)	0.65(4)	0.68(3)	0.69(2)	**0.71(1)**
Average Rank ↓	13.3(13)	13.3(13)	6(7)	10(11)	8.7(9)	11.5(12)	9.2(10)	8(8)	5.3(6)	5.2(5)	3.8(4)	3.2(3)	2.7(2)	**1.3(1)**
Accuracy ↑	0.47(13)	0.37(14)	0.61(5)	0.52(11)	0.58(9)	0.50(12)	0.58(9)	0.61(5)	0.61(5)	60.0(8)	0.64(4)	0.70(2)	0.69(3)	**0.72(1)**

Table 2. Comparison with the state-of-the-art methods on Twitter_LDL dataset.

	PT		AA		SA			CNN-based						
Measures	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS	SA-CPNN	CNNR	DLDL	ACPNN	JCDL	SSDL	E-GCN	Ours
Chebyshev ↓	0.53(13)	0.63(14)	0.28(7)	0.37(11)	0.28(7)	0.37(11)	0.36(10)	0.28(7)	0.26(5)	0.27(6)	0.25(3)	0.25(3)	0.24(2)	**0.22(1)**
Clark ↓	0.85(7)	0.91(14)	**0.58(1)**	0.89(12)	0.86(11)	0.89(12)	0.85(7)	0.84(3)	0.84(3)	0.85(7)	0.83(2)	0.84(3)	0.85(7)	0.84(3)
Canberra ↓	0.77(6)	0.88(14)	**0.41(1)**	0.84(12)	0.79(11)	0.84(12)	0.78(8)	0.76(2)	0.77(6)	0.78(8)	0.76(2)	0.76(2)	0.78(8)	0.76(2)
KL ↓	1.31(12)	1.65(13)	3.89(14)	1.19(10)	0.64(7)	1.19(10)	0.85(9)	0.67(7)	0.54(5)	0.58(6)	0.53(4)	0.51(3)	0.46(2)	**0.44(1)**
Cosine ↑	0.53(13)	0.25(14)	0.82(7)	0.71(11)	0.82(7)	0.71(11)	0.75(10)	0.82(7)	0.83(6)	0.84(5)	0.85(4)	0.86(3)	0.87(2)	**0.89(1)**
Intersection ↑	0.40(13)	0.21(14)	0.66(5)	0.59(9)	0.63(8)	0.57(11)	0.56(12)	0.58(10)	0.65(6)	0.64(7)	0.68(4)	0.69(3)	0.70(2)	**0.72(1)**
Average Rank ↓	10.7(12)	13.8(14)	5.8(6)	10.8(11)	8.5(9)	11.2(13)	9.3(10)	6(7)	5.2(5)	6.5(8)	3.2(3)	2.8(2)	3.8(4)	**1.5(1)**
Accuracy ↑	0.45(13)	0.40(14)	0.73(7)	0.72(9)	0.70(10)	0.57(12)	0.70(10)	0.74(5)	0.73(7)	0.74(5)	0.76(3)	0.77(2)	0.76(3)	**0.78(1)**

Ablation Study

Table 4. Ablation study of loss function on Flickr_LDL dataset.

Measures	$\mathcal{L}_{KL}$	$\mathcal{L}_{KL}+\mathcal{L}_p$	$\mathcal{L}_{KL}+\mathcal{L}_t$	$\mathcal{L}_{KL}+\mathcal{L}_p+\mathcal{L}_t$	$\mathcal{L}_{KL}+\mathcal{L}_{PC}$
Chebyshev ↓	0.239	0.225	0.222	0.218	**0.213**
Clark ↓	0.783	0.779	0.779	0.775	**0.774**
Canberra ↓	0.697	0.689	0.687	**0.682**	0.685
KL ↓	0.435	0.441	0.420	0.414	**0.408**
Cosine ↑	0.843	0.862	0.869	0.870	**0.874**
Intersection ↑	0.678	0.693	0.705	0.703	**0.709**
Accuracy ↑	0.669	0.695	0.700	0.718	**0.721**

Table 5. Ablation study of loss function on Twitter_LDL dataset.

Measures	$\mathcal{L}_{KL}$	$\mathcal{L}_{KL}+\mathcal{L}_p$	$\mathcal{L}_{KL}+\mathcal{L}_t$	$\mathcal{L}_{KL}+\mathcal{L}_p+\mathcal{L}_t$	$\mathcal{L}_{KL}+\mathcal{L}_{PC}$
Chebyshev ↓	0.259	0.240	0.233	0.230	**0.224**
Clark ↓	0.861	0.851	0.848	0.846	**0.842**
Canberra ↓	0.797	0.778	0.775	0.772	**0.764**
KL ↓	0.464	0.476	0.455	0.450	**0.439**
Cosine ↑	0.848	0.870	0.878	0.882	**0.886**
Intersection ↑	0.686	0.706	0.703	0.713	**0.717**
Accuracy ↑	0.744	0.764	0.770	0.779	**0.781**

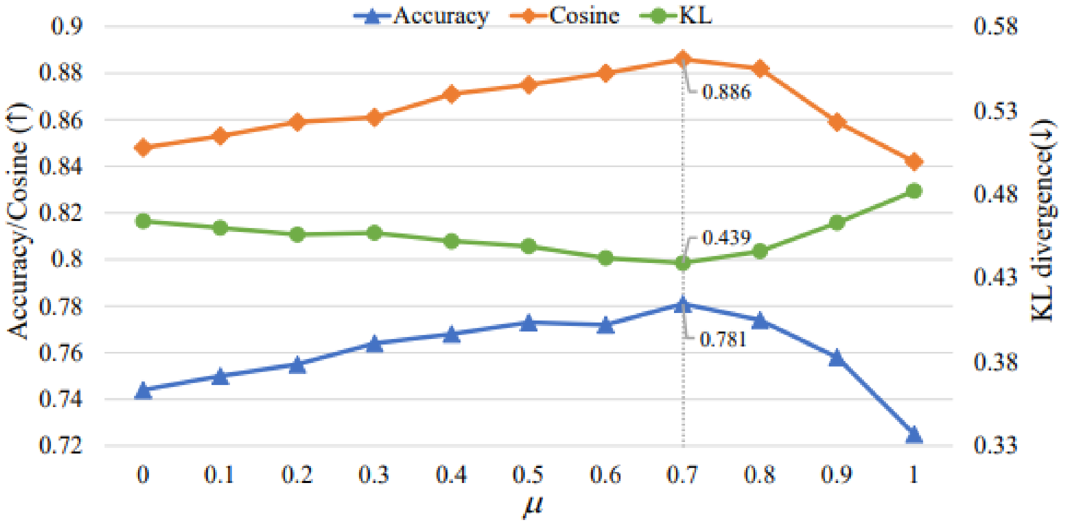


Figure 5. Effect of μ for combined loss on Twitter_LDL dataset. Note that $\mu = 1$ suggests only using PC loss while $\mu = 0$ means implementing KL loss alone.

Visualization

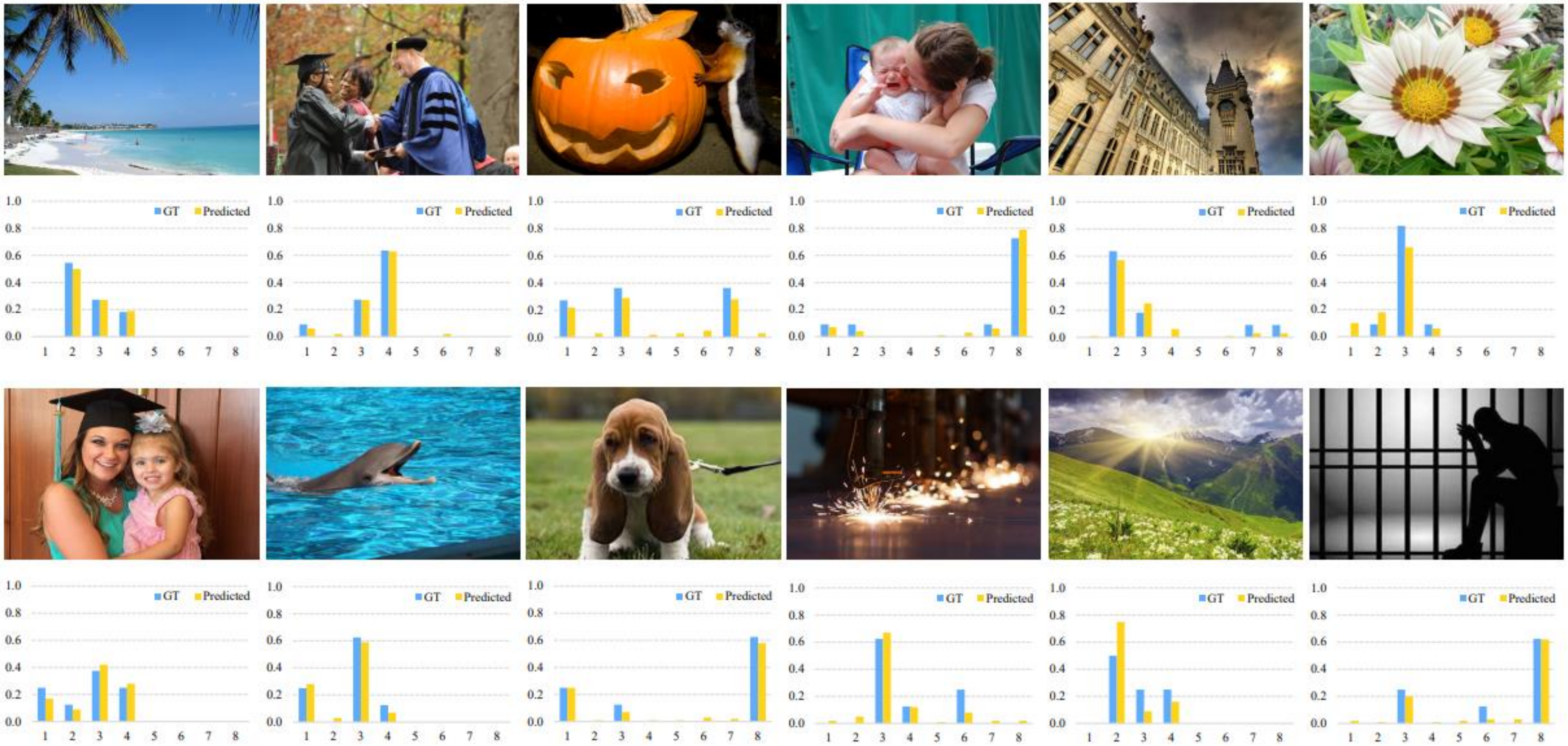


Figure 6. Visualization of the predicted emotion distributions (predicted) and the ground-truth (GT) ones, where images in the first line come from Flickr_LDL dataset and second line the Twitter_LDL. Each number on the horizontal axis corresponds to an emotion category.

Research Directions

- Theoretical study.
  - Learn the Highest Label and Rest Label Description Degrees[8], which focuses on both the majority and the distribution.

- Apply LDL to new tasks
  - Emotion recognition
  - Impression estimation
  - Number counting

- Build new constraints. (tricks?)
  - Find new relationships between different classes.

[8]Wang J. et al. Learn the Highest Label and Rest Label Description Degrees. IJCAI 2021

Thoughts

- At present, LDL has not been systematically carried out in the field of **multimodal.**

- Disadvantages of LDL:
  - Labeling is **subjective**, and in many applications, it is difficult to obtain datasets because building LDL datasets is very time-consuming and labor-intensive.
  - The labeling process may also introduce noise due to subjectivity and other factors, and the effect may not be improved compared with single labeling.

- For practical problems, it needs to be defined according to specific problems depending on whether it’s necessary and can bring improvement.