模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Label Structure Preserving Contrastive Embedding for Multi-Label Learning with Missing Labels

Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li and Yandong Guo

Background

What is multi-label learning?

Partial Multi-Label Learning (PML)

- ✓ PML requires all **positive** labels are annotated.
- ✓ Find the most likely label in the candidate label set.

Multi-Label Learning with Missing Labels (MLML)

- ✓ Complete positive labels are not required in MLML.
- ✓ An unannotated label may be a **true** negative label or **false** negative label.

Settings	label1	label2	label3	label4	label5
Full labels	✓	✓	✗	✗	✗
PML	✓	✓	✓	○	○
MLML	✓	○	○	○	○

Positive: Person, Cake
Missing: Wedding, Plate, Tie

It's difficult and impractical to annotate the full label set!

	Full Labels	PML	MLML
Person	✓	✓	✓
Car	✓	✓	✓
Backpack	✓	✓	○
Dog	✓	✓	○
Chair	×	✓	○

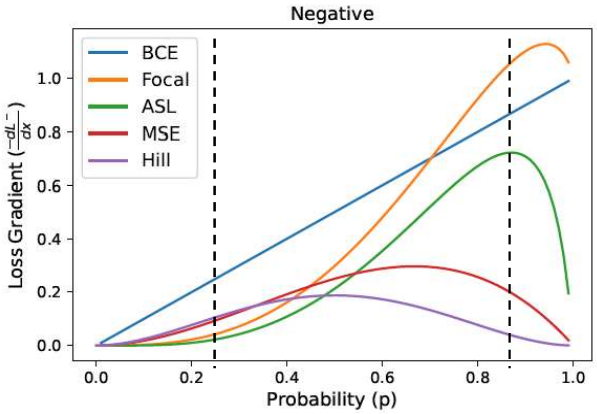
Fig. 1: ✓, , ○ means a label is present, absent, unknown, and falsely marked label is in red.

Background

How to alleviate the effect of false negatives due to miss-labeling?

Loss Function

Due to the positive-negative imbalance, loss function needs to put less weight on **negative** instances, especially on **FN** ones



$$\mathcal{L} = -\sum_{i=1}^{K} \left(y_i L_i^+ + (1 - y_i) L_i^- \right) \quad \begin{cases} L_{BCE}^+ = \log(p) \\ L_{BCE}^- = \log(1-p) \end{cases}$$

$$\begin{cases} L_{Focal}^+ = \alpha_+ (1-p)^\gamma \log(p) \\ L_{Focal}^- = \alpha_- p^\gamma \log(1-p) \end{cases} \quad \begin{cases} L_{ASL}^+ = (1-p_m)^{\gamma_+} \log(p_m) \\ L_{ASL}^- = p_m^{\gamma_-} \log(1-p_m) \end{cases}$$

Discard very easy negatives, down-weight low probability negatives, and reject hard negatives

Noisy Label

Missing label can be regarded as a branch of label noise problem, which means that we can find the **noise transition matrix T** that characterizes the probabilities of a training example being wrongly annotated. (Find anchor point? KNN?...)

Network Architecture

Use transformer-based model to locate multiple regions of interests.

Motivation

Correlations between labels and between instances can be very helpful in finding missing labels.

Fish
Seaweed
Ocean(missing)
…
…

Fish(missing)
Seaweed
Ocean
…
…

How to evaluate similarities between samples? Directly use Contrastive Learning?

Hard to separate the positive and negative instances due to label missing, especially when the number of categories goes large.

Low-rank local label dependency

The rank of the label sub-matrix for samples that share the **same** label should be small.

High-rank global label dependency

The rank of the label matrix for all samples should be large because of label diversity.

Method

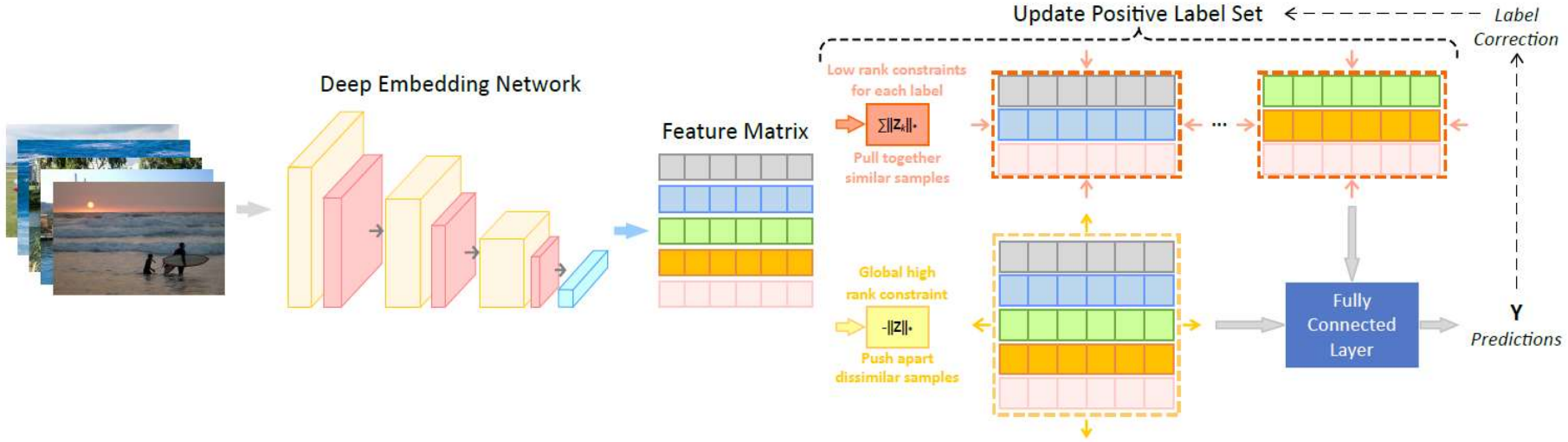
Let $\mathbf{X} \in R^{N\times W\times H\times 3}$ denote a batch of images, where N is the batch size, H and W are the height and width of the images. For the i-th image $\mathbf{x}_i \in \mathbf{X}$, we denote $\mathbf{y}_i^o \in \{-1,+1\}^{|C|}$ the observed label vector, and $\mathbf{y}_i \in \{-1,+1\}^{|C|}$ the ground truth label vector, where C is the size of label set.

Contrastive loss: $\mathcal{L}_{CL} := \sum_{k=1}^{C} \|\mathbb{E}(\mathbf{X}_k;\theta)\|_* - \|\mathbb{E}(\mathbf{X};\theta)\|_*$

Increase the intra-class **similarity**

Increase inter-class **separation**

$\mathbf{Z} = \mathbb{E}(\mathbf{X};\theta)$
The deep embedding of $N \times D$ dimension of input instance sub-matrix X



Method

Update positive label set

The positive label set with Label Correction (LaCo) can be represented as follows

$$LaCo(y_{ij}^o) = \begin{cases} +1 \ , & f\left(\mathbf{Z}_i\right)_j \geq \delta \\ y_{ij}^o \ , & f\left(\mathbf{Z}_i\right)_j < \delta \end{cases}$$

If the prediction is larger than threshold, then we can think of it as a **false negative**(label missing) sample and change its label.

CLML loss can be expressed as:

$$\mathcal{L}_{CLML} = \sum_{k=1}^{C} || \left\{ \mathbf{Z}_i | \widetilde{y}_{ik} = +1, \mathbf{Z}_i \in \mathbf{Z} \right\} ||_* - \|\mathbf{Z}\|_*$$

$$\widetilde{y}_{ik} = \begin{cases} y_{ij}^o \ , & N_e < N_E \\ LaCo(y_{ij}^o) \ , & N_e \geq N_E \end{cases}$$

Where N_E indicates that the positive label set starts updating.

The final loss is: $\min_{\theta,\phi} \mathcal{L}_{classification}\left(\mathbf{X}, \mathbf{Y}^o, \theta, \phi\right) + \lambda \cdot \mathcal{L}_{CLML}\left(\mathbf{X}, \mathbf{Y}^o, \theta\right)$

Method

Algorithm 1 Label Structure Preserving Contrastive Embedding

Input: Training data matrix $\mathbf{X}$, label matrix $\mathbf{Y}^o$, deep embedding network $\mathbf{Z} = \mathbb{E}(\mathbf{X}, \theta)$ and deep multi-label classifier $f(\mathbf{Z}, \phi)$, trade-off parameter λ.

Output: The well trained deep model $\mathbb{E}(\cdot, \theta)$ and $f(\cdot, \phi)$

```
for N_e = {1, ··· , N_epoch} do
    for each minibatch X_b and Y_b^o do
        Z_b = E(X_b, θ)
        Correct the false negative labels in Y_b^o according to Eq. (6)
        Calculate contrastive loss L_CL(Z_b, Y_b^o) according to Eq. (7)
        Calculate total loss according to Eq. (10)
        Backpropagation
    end for
end for
```

$$LaCo(y_{ij}^o) = \begin{cases} +1 \ , & f(\mathbf{Z}_i)_j \geq \delta \\ y_{ij}^o \ , & f(\mathbf{Z}_i)_j < \delta \end{cases}$$

$$\mathcal{L}_{CLML} = \sum_{k=1}^{C} || \{\mathbf{Z}_i | \widetilde{y}_{ik} = +1, \mathbf{Z}_i \in \mathbf{Z}\} ||_* - ||\mathbf{Z}||_*$$

Experiment

TABLE I: Specific data from different datasets

	samples	classes	labels	avg.label/img
COCO-full labels	82,081	80	241,035	2.9
COCO-75% labels left	82,081	80	181,422	2.2
COCO-40% labels left	82,081	80	96,251	1.2
COCO-single label	82,081	80	82,081	1.0
NUS-full labels	119,103	81	289,460	2.4
NUS-single label	119,103	81	119,103	1.0
VOC-full labels	5,717	20	8,331	1.4
VOC-single label	5,717	20	5,717	1.0

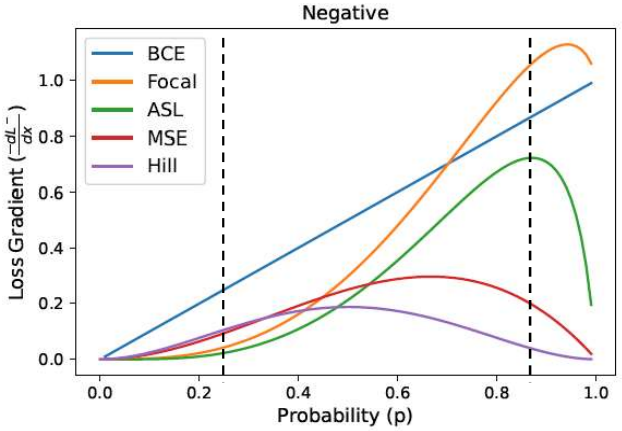
Experiment setting

Experiment

TABLE II: Compared results on COCO dataset with varied missing label ratios

Method		BCE (full labels)	BCE	BCE+CLML	Focal [41]	Focal+CLML	Hill [43]	Hill+CLML	SPLC [43]	SPLC+CLML
75% labels left	mAP ↑	80.3	76.8	**78.0**	77.0	**78.3**	78.8	**79.6**	78.4	**80.4**
	CP↑	80.8	85.1	**86.2**	83.8	**86.0**	**73.6**	72.8	72.6	**75.6**
	CR↑	70.3	58.1	**58.7**	59.4	**61.0**	74.4	**76.3**	**75.1**	74.6
	CF1↑	74.9	67.7	**68.5**	68.4	**69.7**	73.6	**74.1**	73.2	**74.8**
	OP↑	84.3	90.1	**90.9**	88.6	**89.1**	**76.4**	74.6	74.0	**79.1**
	OR↑	74.2	58.7	**59.3**	59.8	**61.2**	78.3	**80.3**	**79.3**	78.0
	OF1↑	78.9	71.1	**71.8**	71.4	**72.6**	77.3	77.3	76.6	**78.5**
40% labels left	mAP↑	-	70.5	**71.5**	71.7	**73.3**	75.2	**76.4**	75.7	**76.5**
	CP↑	-	89.2	**89.5**	88.9	**89.2**	80.4	**81.2**	**81.6**	79.6
	CR↑	-	34.4	**36.0**	37.0	**40.3**	61.4	**62.6**	60.7	**65.8**
	CF1↑	-	45.8	**47.8**	48.7	**52.1**	68.6	**69.6**	67.9	**70.8**
	OP↑	-	**94.1**	93.7	**93.7**	93.3	85.5	**96.5**	**87.7**	83.6
	OR↑	-	25.7	**27.3**	28.8	**31.2**	63.8	**64.6**	63.0	**69.5**
	OF1↑	-	40.4	**42.3**	44.0	**47.0**	73.1	**74.0**	73.3	**75.9**
single label	mAP↑	-	68.6	**69.5**	70.2	**71.8**	73.2	**74.0**	73.2	**74.0**
	CP↑	-	88.6	**89.1**	88.2	**88.9**	79.7	**83.0**	**83.8**	80.9
	CR↑	-	33.0	**33.5**	36.0	**37.4**	**58.0**	55.7	53.1	**58.7**
	CF1↑	-	43.8	**44.2**	47.0	**48.6**	**65.5**	64.2	61.6	**65.5**
	OP↑	-	93.9	**94.8**	93.4	**93.9**	85.3	**88.7**	**90.1**	86.4
	OR↑	-	23.6	**24.5**	26.6	**28.3**	**58.7**	55.0	53.8	**60.5**
	OF1↑	-	37.7	**38.9**	41.4	**43.5**	**69.5**	67.8	67.4	**71.2**

Aside



Simple and Robust Loss Design for Multi-Label Learning with Missing Labels

Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li and Yandong Guo

"Hill" loss can be calculated by re-weighting MSE loss

$$\mathcal{L}^-_{Hill} = -w(p) \times MSE = -(\lambda - p)p^2. \Rightarrow p_m = \max(p - m, 0)$$

Probability shift

Self-Paced Loss Correction (SPLC)

$$\begin{cases} L^+ = \log(p) \\ L^- = \mathbb{I}(p \leq \tau)\log(1-p) + (1 - \mathbb{I}(p \leq \tau))\log(p) \end{cases}$$

Experiment

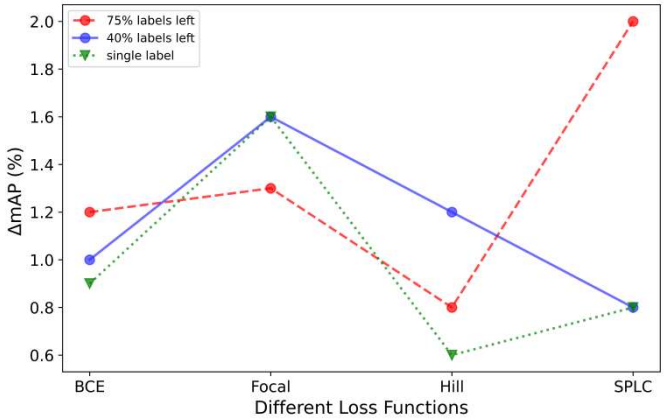


Fig. 4: The increase amplitude of CLML to different loss functions on training set with different ratios of missing labels.

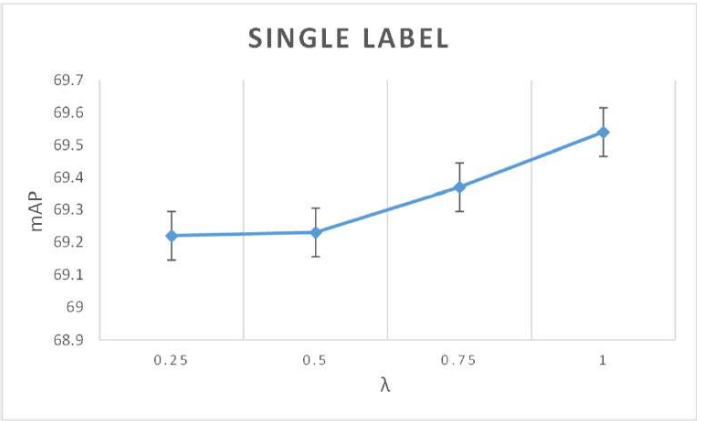


Fig. 5: In the case of single label, the influence of different λ on model performance.

Experiment

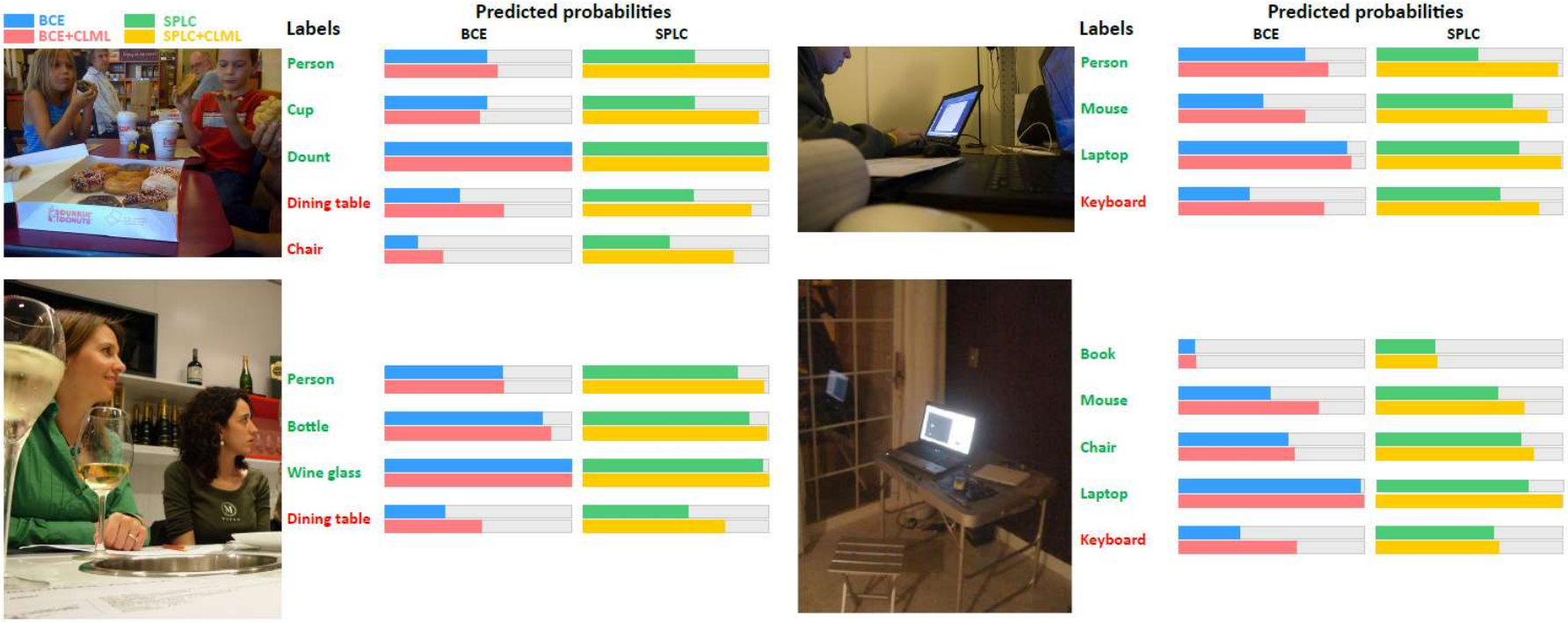


Fig. 6: On the COCO validation set, we compare BCE loss and SPLC loss with and without CLML. Positive labels and missing labels are represented by green and red, respectively. For the predicted probabilities, blue and pink represent BCE loss with and without CLML, respectively. Green and orange represent SPLC loss with and without CLML, respectively. The results show that loss with CLML can better predict missing labels via label correlation, such as the strong association between food and dining table, as well as the strong correlation between laptop, mouse, and keyboard.

Experiment

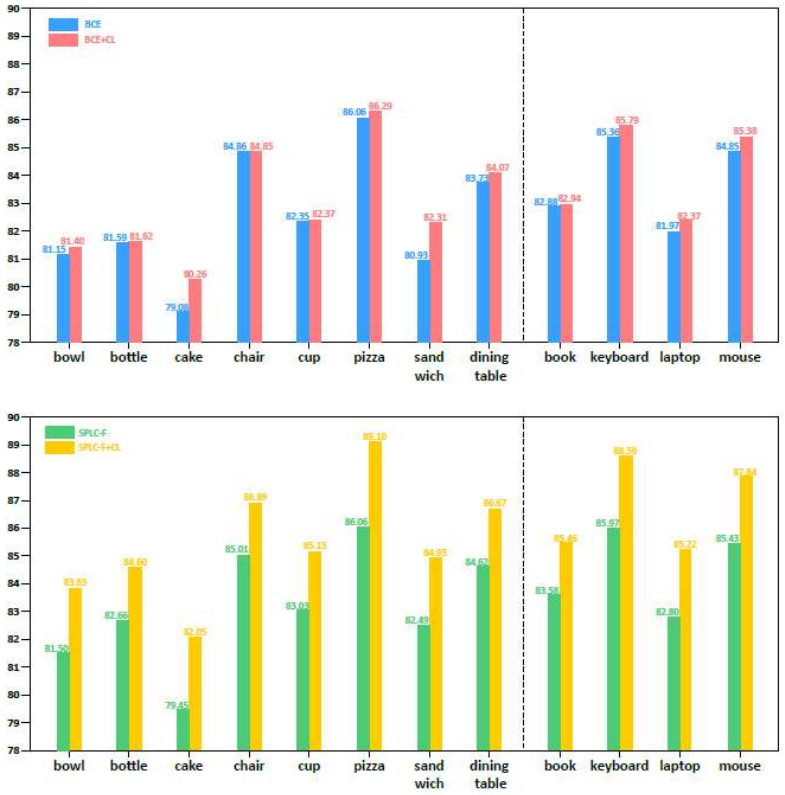


Fig. 7: AP scores of four loss functions in different categories. The classification performance of the BCE loss with and without CLML is on the upper side, and the classification performance of the SPLC loss with and without CLML is on the lower side.

Experiment

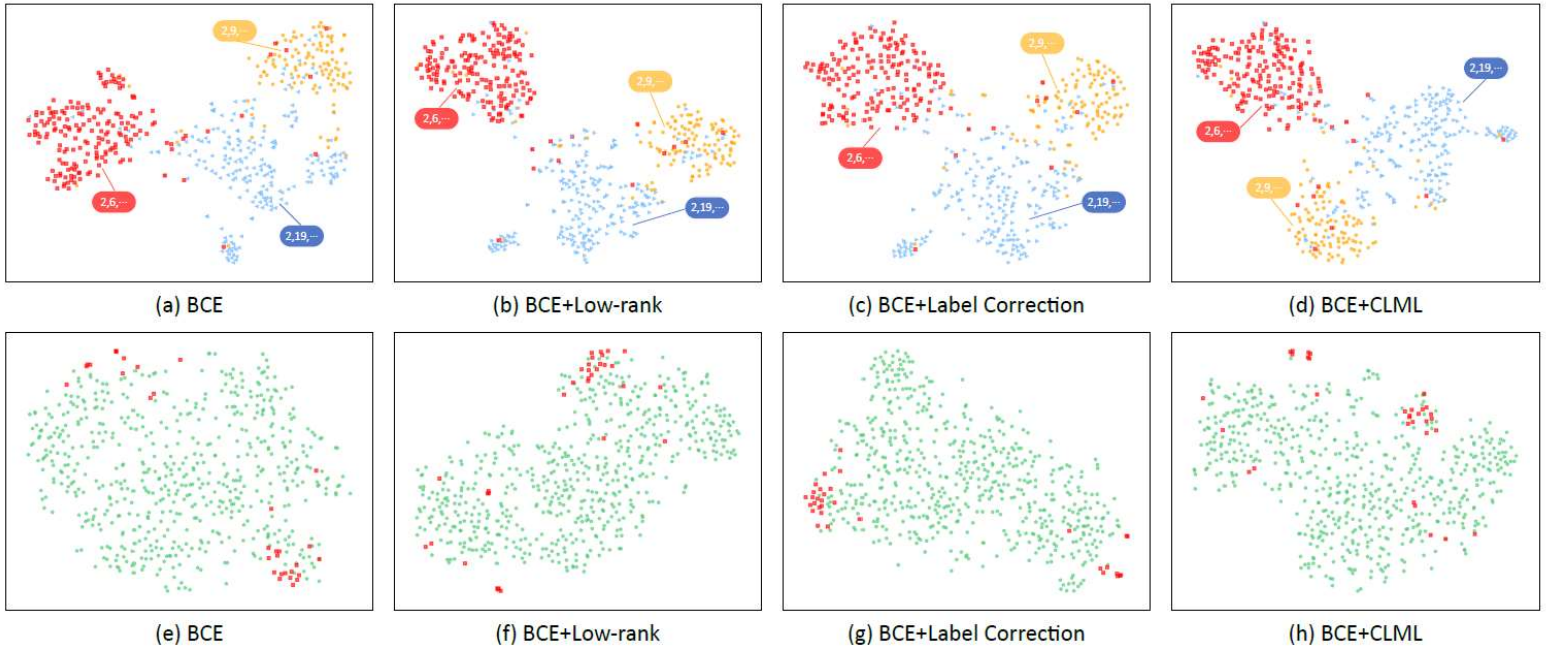


Fig. 8: t-SNE visualization for the learned embedding features of the multi-label images on VOC 2012 test set. Subfigures (a) to (h) are generated by ResNet101 and its corresponding loss function respectively. For subfigures (a) to (d), different categories are indicated by different colors, shapes, and numbers, respectively. For subfigures (e) to (h), green represents the true positive labels, and red represents the missing labels, i.e., the false negative labels.

Experiment

TABLE IV: Compared result on VOC and NUS datasets

Method	VOC-single label			NUS-single label		
	mAP↑	CF1↑	OF1↑	mAP↑	CF1↑	OF1↑
BCE(full labels)	89.0	83.3	85.6	60.6	59.1	73.6
BCE	85.6	77.4	78.4	51.7	30.9	33.3
BCE+CLML	**87.2**	**79.3**	**80.3**	**53.0**	**31.1**	**34.6**
Focal [41]	86.8	78.2	79.1	53.6	34.2	35.0
Focal+CLML	**87.5**	**79.8**	**80.3**	**54.7**	**35.0**	**36.5**
Hill [43]	87.8	81.1	**83.8**	55.0	**54.1**	**68.6**
Hill+CLML	**88.1**	**81.5**	82.0	**55.4**	51.2	65.6
SPLC [43]	**88.1**	80.2	**83.0**	55.2	**52.4**	**70.6**
SPLC+CLML	88.0	**81.5**	82.4	**55.3**	48.6	61.0

TABLE V: Ablation study of different loss functions. The best value is in bold and the second best value is underlined

Method	VOC-single label						
	mAP↑	CP↑	CR↑	CF1↑	OP↑	OR↑	OF1↑
BCE only	85.6	87.8	71.6	77.4	91.1	68.7	78.4
BCE with Low-rank	<u>87.0</u>	**91.2**	70.8	78.3	**93.5**	69.7	79.9
BCE with Label Correction	86.8	89.5	<u>72.2</u>	<u>78.5</u>	92.1	**72.1**	**80.9**
BCE with CLML	**87.2**	<u>90.4</u>	**72.4**	**79.3**	<u>92.2</u>	<u>71.1</u>	<u>80.3</u>

Thanks