南京航空航天大學
Nanjing University of Aeronautics and Astronautics

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

Exploring Structured Semantic Prior for Multi Label Recognition with Incomplete Labels

Zixuan Ding[1,4*] Ao Wang[2,3,4*] Hui Chen[2,3,†] Qiang Zhang[1]
Pengzhang Liu[5] Yongjun Bao[5] Weipeng Yan[5] Jungong Han[6,7]
[1]Xidian University [2]Tsinghua University [3]BNRist
[4]Hangzhou Zhuoxi Institute of Brain and Intelligence [5]JD.com
[6]Department of Computer Science, the University of Sheffield, UK
[7]Centre for Machine Intelligence, the University of Sheffield, UK
dingzixuan@stu.xidian.edu.cn wa22@mails.tsinghua.edu.cn qzhang@xidian.edu.cn
{jichenhui2012,jungonghan77}@gmail.com {Paul.yan, baoyongjun, liupengzhang}@jd.com

CVPR 2023

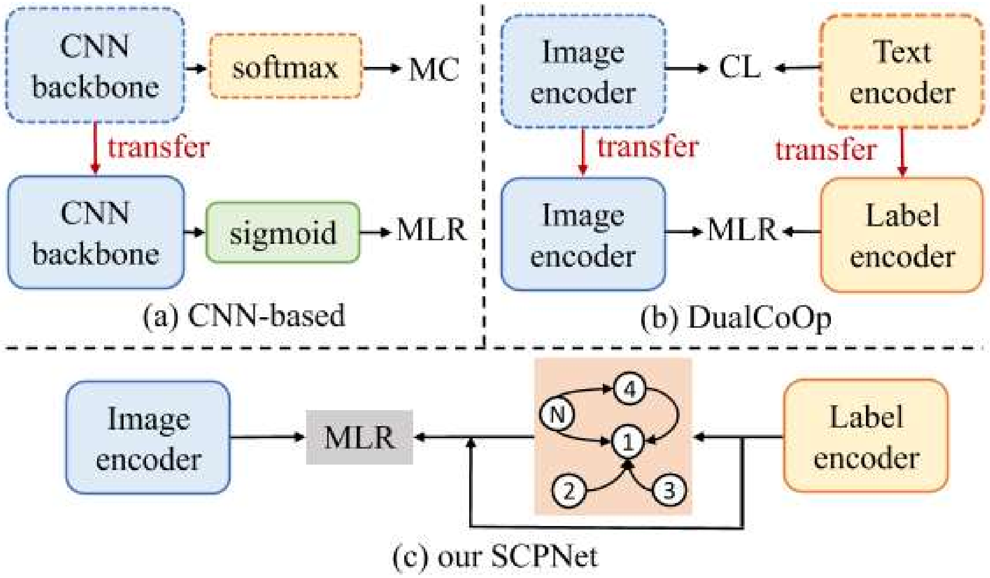
Introduction

- Multi-label recognition (MLR) aims to recognize various semantic labels for an image, which has a wide range of practical applications. However, collecting high-quality full annotations becomes very challenging. Recently, researchers propose to simplify the full label setting to a partial label setting, which annotate a few labels for each training image. Another extreme setting is observed with only one single positive label.

- These settings can be unified into a common issue of incomplete labels, which eases the burden of the full annotation and significantly reduces annotation cost.

Motivation

- However, the incomplete labels setting introduces a dilemma of poor supervisions, resulting in severe performance drops for MLR.
- Existing methods strive to regain supervisions from missing labels by exploring the image-to-label correspondence. Such methods eliminate priors about the correspondence between images and labels although it is necessary and unavoidable.



SCPNet

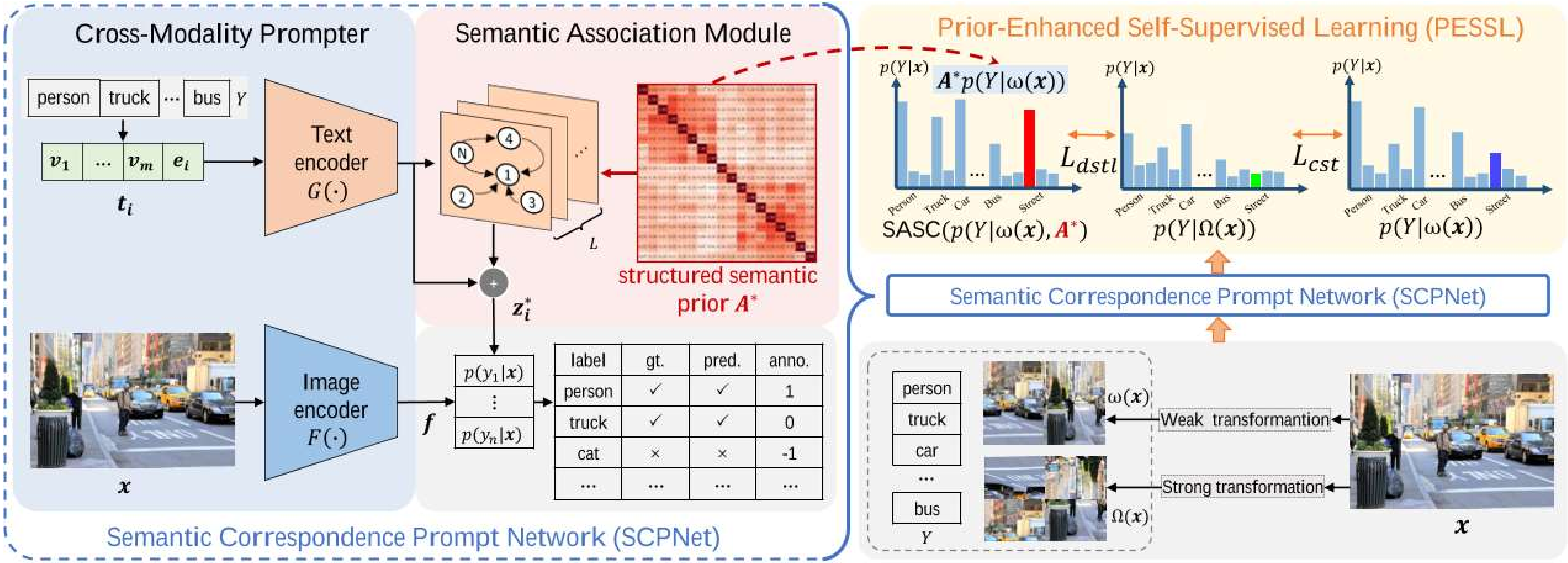


Figure 2. An overview of the proposed method. We design a semantic correspondence prompt network to explore the structured semantic prior for MIR with incomplete labels. A prior-enhanced self-supervised learning strategy is used to enhance such exploration.

Methodology

Structured Prior Prompter

- label prior correlation : $a_{ij} = \mathrm{sim}(\bar{z}_i, \bar{z}_j)$
- label correspondence graph G : $$a^*_{ij} = \frac{\mathrm{I}[\bar{a}_{ij} \neq 0] \exp(\bar{a}_{ij}/\tau')}{\sum_j \mathrm{I}[\bar{a}_{ij} \neq 0] \exp(\bar{a}_{ij}/\tau')}$$

Semantic Correspondence Prompt Network

- **Cross-modality prompter (CMP) :**
- feature extraction process : $f = F(x), z_i = G(t_i).$
- **Semantic association module (SAM): structured semantic prior A***
- GCN layer : $H^{l+1} = \rho(A^* H^l W^l)$

Methodology

Prior-Enhanced Self-Supervised Learning

- **Structure-aware semantic calibration**
- Weighted likelihood by the correlated neighboring labels :

$$p^*(y_i|\boldsymbol{x}) = \sum\nolimits_{y_j \in \mathcal{N}(y_i)} w(i,j) \times p(y_j|\boldsymbol{x})$$

- Formulate process : $\mathrm{SASC}(\boldsymbol{p}(Y|\boldsymbol{x}), \mathbb{W}) = \mathbb{W}\boldsymbol{p}(Y|\boldsymbol{x})$
- **Prior-enhanced learning**
- Self-supervised consistency loss: $\mathcal{L}_{cst} = -\sum\nolimits_{c \in \mathcal{O}(\boldsymbol{x})}^{Y} \log p(c|\Omega(\boldsymbol{x})) - \sum\nolimits_{c \notin \mathcal{O}(\boldsymbol{x})}^{Y} \log(1 - p(c|\Omega(\boldsymbol{x})))$
- Self-distillation loss : $\mathcal{L}_{dstl} = -\sum_{c}^{Y} \left(q_c^w \log \frac{q_c^s}{q_c^w} + (1 - q_c^w) \log \frac{1 - q_c^s}{1 - q_c^w} \right)$
- overall objective : $\mathcal{L}_{pessl} = \lambda_{cst}\mathcal{L}_{cst} + \lambda_{dstl}\mathcal{L}_{dstl}$

Network Optimization $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{pessl}$

Experimental results

- Single positive label

Method	LargeLoss setup [16]					SPLC setup [32]				
	COCO	VOC	NUS	CUB	Avg.	COCO	VOC	NUS	CUB	Avg.
LSAN [8]	69.2	86.7	50.5	17.9	56.1	70.5	87.2	52.5	18.9	57.3
ROLE [8]	69.0	88.2	51.0	16.8	56.3	70.9	89.0	50.6	20.4	57.7
LargeLoss [16]	71.6	89.3	49.6	21.8	58.1	-	-	-	-	-
Hill [32]	-	-	-	-	-	73.2	87.8	55.0	18.8	58.7
SPLC [32]	72.0	87.7	49.8	18.0	56.9	73.2	88.1	55.2	20.0	59.1
SCPNet (ours)	**75.4**	**90.1**	**55.7**	**25.4**	**61.7**	**76.4**	**91.2**	**62.0**	**25.7**	**63.8**

- Partial label

Datasets	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
COCO	SSGRL [5]	62.5	70.5	73.2	74.5	76.3	76.5	77.1	77.9	78.4	74.1
	GCN-ML [6]	63.8	70.9	72.8	74.0	76.7	77.1	77.3	78.3	78.6	74.4
	SST [4]	68.1	73.5	75.9	77.3	78.1	78.9	79.2	79.6	79.9	76.7
	SARB [21]	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
	DualCoOp [26]	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9
	SCPNet (ours)*	**80.3**	**82.2**	**82.8**	83.4	83.8	83.9	84.0	84.1	84.2	83.2
	SCPNet (ours)	79.1	82.1	**82.8**	**83.9**	**84.5**	**84.9**	**85.4**	**85.7**	**85.9**	**83.8**
VOC2007	SSGRL [5]	77.7	87.6	89.9	90.7	91.4	91.8	91.9	92.2	92.2	89.5
	GCN-ML [6]	74.5	87.4	89.7	90.7	91.0	91.3	91.5	91.8	92.0	88.9
	SST [4]	81.5	89.0	90.3	91.0	91.6	92.0	92.5	92.6	92.7	90.4
	SARB [21]	83.5	88.6	90.7	91.4	91.9	92.2	92.6	92.8	92.9	90.7
	DualCoOp [26]	90.3	92.2	92.8	93.3	93.6	93.9	94.0	94.1	94.2	93.2
	SCPNet (ours)	**91.1**	**92.8**	**93.5**	**93.6**	**93.8**	**94.0**	**94.1**	**94.2**	**94.3**	**93.5**
VG-200	SSGRL [5]	34.6	37.3	39.2	40.1	40.4	41.0	41.3	41.6	42.1	39.7
	GCN-ML [6]	32.0	37.8	38.8	39.1	39.6	40.0	41.9	42.3	42.5	39.3
	SST [4]	38.8	39.4	41.1	41.8	42.7	42.9	43.0	43.2	43.5	41.8
	SARB [21]	41.4	44.0	44.8	45.5	46.6	47.5	47.8	48.0	48.2	46.0
	SCPNet (ours)	**43.8**	**46.4**	**48.2**	**49.6**	**50.4**	**50.9**	**51.3**	**51.6**	**52.0**	**49.4**

For both setups, our method can significantly outperform existing methods on all benchmark datasets, achieving state-of-the-art performance

Ablation study

- Introduce a model that directly employ L_{cls} to optimize a ResNet-based MLR model, as the baseline.
- Each component can obtain consistent performance improvement in all datasets.

Table 3. Effect of different modules in the proposed SCPNet method for both the single positive label setting and the partial label setting (%). An average of all metrics is also reported.

Model	CMP	SAM	PESSL		Single Positive Label				Partial Label			Avg.
			$\mathcal{L}_{cst}$	$\mathcal{L}_{dstl}$	COCO	VOC	NUS	CUB	COCO	VOC2007	VG-200	
Baseline					73.18	88.07	55.18	19.99	77.41	88.32	46.39	64.08
SCPNet	✓				74.36	88.46	60.66	21.42	80.90	89.16	47.55	66.07
	✓	✓			75.12	89.09	61.08	21.66	82.12	90.16	48.11	66.76
	✓	✓	✓		75.70	90.92	61.75	23.67	82.85	92.50	48.70	68.01
	✓	✓		✓	75.84	90.92	61.56	24.51	83.35	93.21	48.83	68.32
	✓	✓	✓	✓	**76.42**	**91.16**	**62.04**	**25.71**	**83.76**	**93.49**	**49.36**	**68.85**

Visualization Results

- Visualize the structured semantic prior and the label representation in the label feature space.

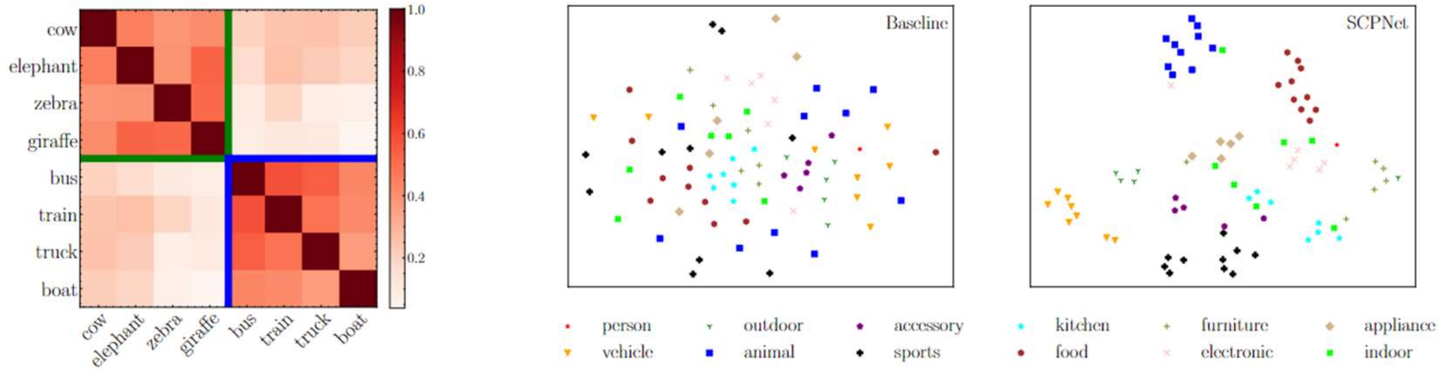


Figure 3. The structured semantic prior (left) and the learnt label representation (middle: in the baseline, right: in our SCPNet).

- Visualize the precision variations to verify the effect of the proposed structured semantic prior.

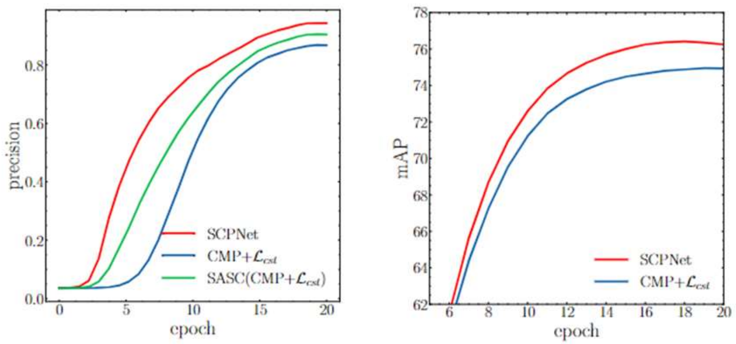


Figure 4. The precision on the training set (left) and the mAP on the test set (right).

Model analysis

- **Correlation graph construction**

Table 4. Analysis on the correlation graph.

SAM	PESSL	mAP (%)
Static	Static	**76.42**
	Dynamic	76.05
	No	75.83
Dynamic	Static	76.08
	Dynamic	75.84

- **Prior knowledge extraction**

Table 5. Analysis on the prior extraction (%).

Prior	Dynamic	Image	Glove	BERT	CLIP
mAP	75.84	75.67	76.15	76.16	**76.42**

- **Generalization on the CNN-based architecture**

Table 6. Prior for MLR models with the ImageNet-based ResNet.

Image Encoder	Label Encoder	mAP (%)
ResNet	sigmoid	73.18
ResNet	Ours	74.72
Ours	Ours	**76.42**

- **Analysis on hyper-parameters**

Table 7. Analysis on the number of GCN Layer, *i.e.*, L (%).

L	2	3	4
mAP	75.88	**76.42**	76.22

Table 8. Analysis on λ_{cst} and λ_{dstl} (%).

λ_{cst}	0	1/16	1/8	1/4	1/8		
λ_{dstl}	0				1/8	2/8	3/8
mAP	75.12	75.56	**75.70**	75.13	**76.42**	76.40	76.17

THANKS