Semantic-Aware Dual Contrastive Learning for Multi-label Image Classification

Leilei Ma[1], Dengdi Sun[2], Lei Wang[3], Haifeng Zhao[1,4;✉] and Bin Luo[1]

[1]School of Computer Science and Technology, Anhui University, China

[2]School of Artificial Intelligence, Anhui University, China

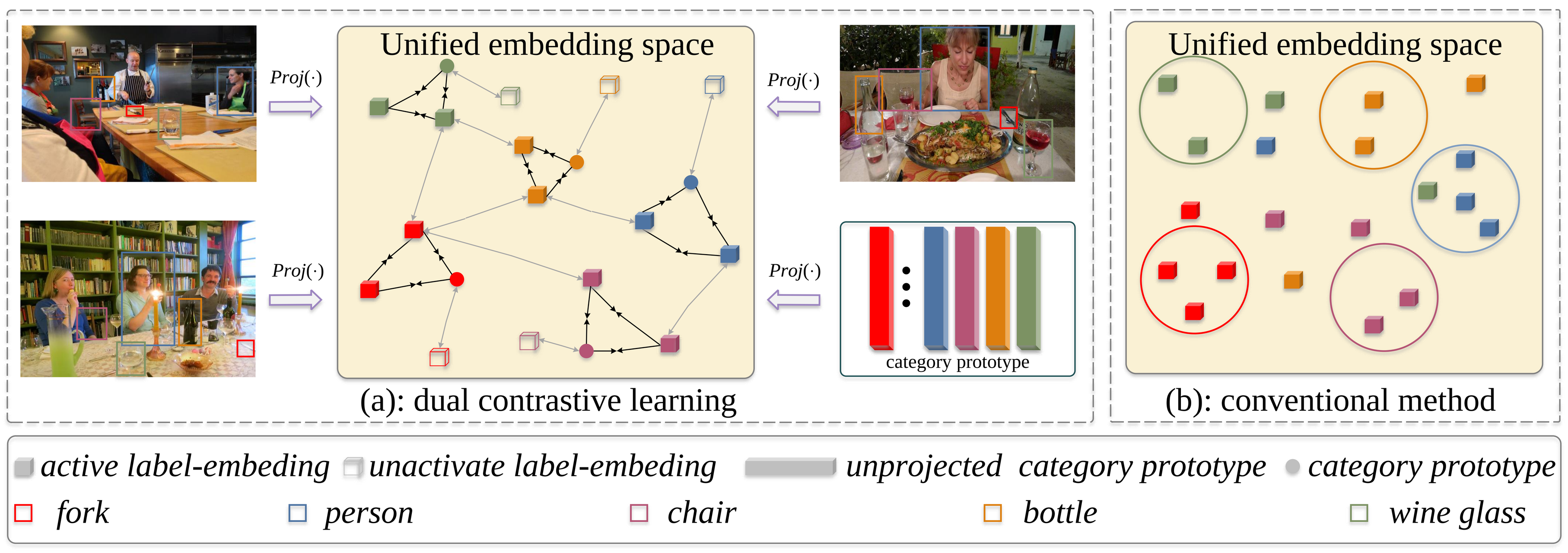
[3]School of Computer Science and Engineering, Nanjing University of Science and Technology, China

[4]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China

Highlights

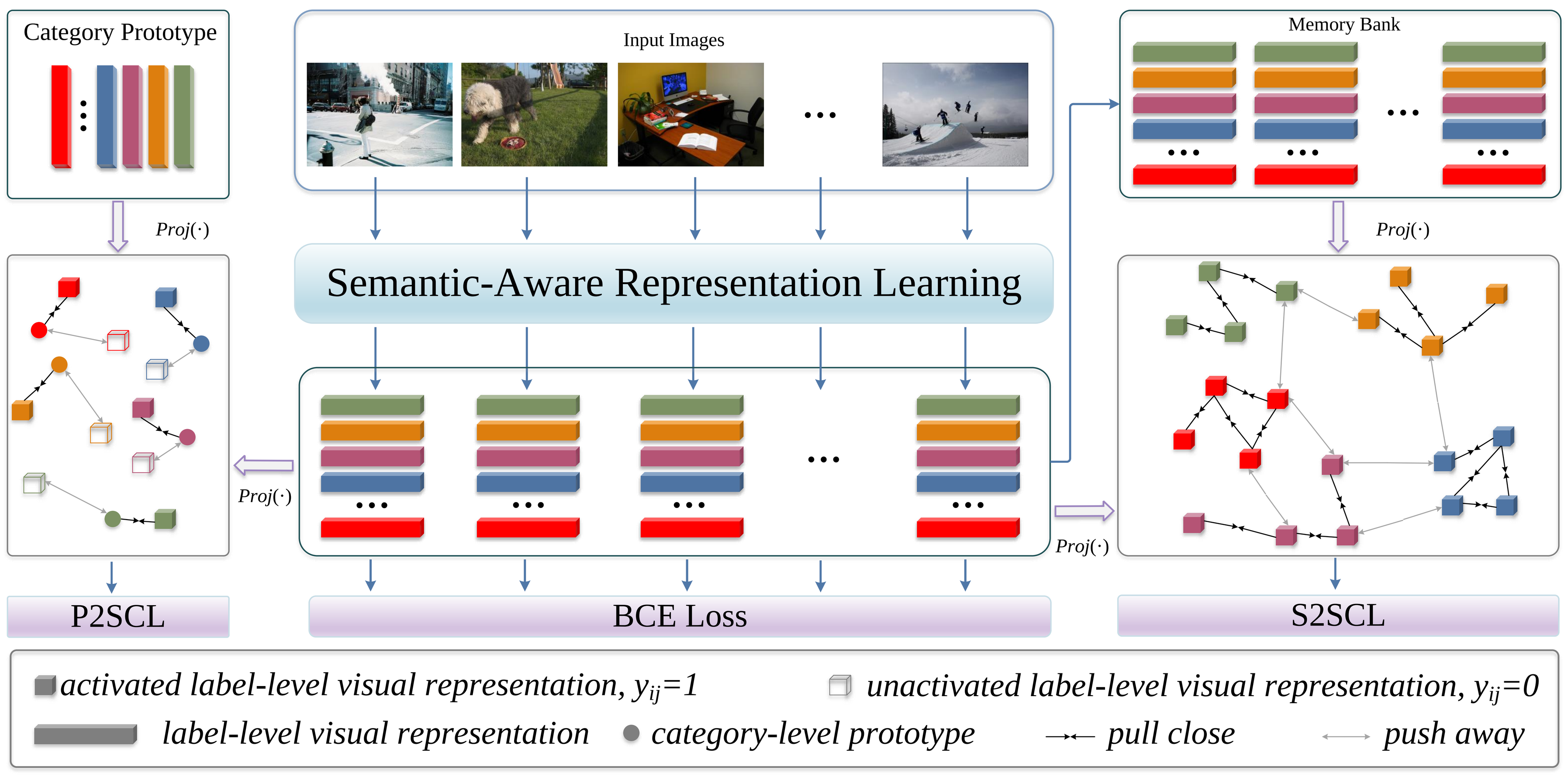
- We propose a novel Semantic-Aware Dual Contrastive Learning framework named SADCL for multi-label image classification, effectively learning more discriminative visual representation.
- Compared with class activation maps (CAM), we leverage Semantic-Aware Representation Learning to accurately and easily locate the label-related image regions.
- Experiments on five challenging large-scale public datasets (MS-COCO, PASCAL VOC 2007&2012, NUS-WIDE, and Visual Genome) show that our proposed method is effective and outperforms the state-of-the-art methods.

Problems and Motivations



- Common methods fail to localize the semantic region of interest in the image, or the localized object region lacks discrimination and contains potential noise.
- Existing methods consider only inter-category relationships (intra-image), ignoring intra-category relationships (cross-image).

Framework and Method



$$\mathcal{L}_{S2S}^{ij} = \sum_{x_{ij(p)} \in \mathcal{P}_{ss}(i,j)} \log \frac{\exp(x_{ij} \cdot x_{ij(p)}/\tau)}{\sum_{x_a \in \mathcal{A}(i,j)} \exp(x_{ij} \cdot x_a/\tau))}, \mathcal{L}_{S2S} = \sum_{x_{ij} \in \mathcal{I}} \frac{-1}{|\mathcal{P}_{ss}(i,j)|} \mathcal{L}_{S2S}^{ij}$$

Sample-to-Sample Contrastive Learning

$$\mathcal{L}_{P2S} = -\sum_{j \in C} \log \frac{\sum_{x_{ij(p)} \in \mathcal{P}_{sp}(i,j)} \exp(c_j \cdot x_{ij(p)}/\tau)}{\sum_{x_{ij} \in \mathcal{P}_{sp}(i,j) \cup \mathcal{N}_{sp}(i,j)} \exp(c_j \cdot x_{ij}/\tau)}$$

Prototype-to-Sample Contrastive Learning

- We model the context relationship among multiple objects and scenes at the front end of the framework, and generate an initial label-level visual representation with abundant semantic information through transformer autoencoder with multi-head attention mechanism.
- Our method aims to maximize the inter-class distance and minimize the intra-class distance of visual representations in the unified embedding space.
- Sample-to-sample contrastive learning considers only activated label-level visual representations. We propose a prototype-based contrastive learning loss to fully exploit this unactivated label-level visual representation information.

Experiments

Methods	(R_{train}, R_{test})	mAP	All CP	All CR	All CF1	All OP	All OR	All OF1	Top 3 CP	Top 3 CR	Top 3 CF1	Top 3 OP	Top 3 OR	Top 3 OF1
CNN-RNN [25]	$(-,-)$	61.2	-	-	-	-	-	-	66.0	55.6	60.4	69.2	66.4	67.8
RNN-Att [26]	$(-,-)$	-	-	-	-	-	-	-	79.1	58.7	67.4	84.0	63.0	72.0
ResNet101*[12]	(448, 448)	81.5	82.1	71.2	76.0	84.6	75.4	79.7	85.9	62.9	71.6	89.6	66.1	76.1
MLGCN [5]	(448, 448)	83.0	85.1	72.0	78.0	85.8	75.4	80.3	89.2	64.1	74.6	90.5	66.5	76.7
MS-CMA [34]	(448, 448)	83.8	82.9	74.4	78.4	84.4	77.9	81.0	88.2	65.0	74.9	90.2	67.4	77.1
P-GCN [6]	(448, 448)	83.2	84.9	72.7	78.3	85.0	76.4	80.5	89.2	64.3	74.8	90.0	66.8	76.7
GM-MLIC [29]	(448, 448)	84.3	87.3	70.8	78.3	88.6	74.8	80.6	90.6	67.3	74.9	94.0	69.8	77.8
MCAR [10]	(448, 448)	83.8	85.0	72.1	78.0	88.0	73.9	80.3	88.1	65.5	75.1	91.0	66.3	76.7
TDRG [35]	(448, 448)	84.6	86.0	73.1	79.0	86.6	76.4	81.2	89.9	64.4	75.0	91.2	67.0	77.2
CCD-R101 [20]	(448, 448)	84.0	87.2	70.9	77.3	88.8	74.6	81.1	89.7	63.9	72.9	92.0	66.5	77.2
MulCon [8]	(448, 448)	84.9	84.0	74.8	79.2	85.6	78.0	81.6	87.8	65.9	75.3	90.5	67.9	77.6
Query2Label [21]	(448, 448)	84.9	84.8	74.5	79.3	86.6	76.9	81.5	78.0	69.1	73.3	80.7	70.8	75.4
CPSD [30]	(448, 448)	84.9	88.4	71.7	79.2	89.3	74.8	81.4	-	-	-	-	-	-
Ours(SADCL)	(448, 448)	85.6	84.6	76.0	79.8	86.0	78.5	82.1	88.9	66.6	74.9	91.0	68.3	78.0
ADDGCN [33]	(448, 576)	85.2	84.7	75.9	80.1	84.9	79.4	82.0	88.8	66.2	75.8	90.3	68.5	77.9
SSGRL [2]	(576, 576)	83.8	89.9	68.5	76.8	91.3	70.8	79.7	91.9	62.5	72.7	93.8	64.1	76.2
AdaHGNN [28]	(576, 576)	85.0	-	-	79.9	-	-	81.8	-	-	75.5	-	-	77.6
C-Tran [16]	(576, 576)	85.1	86.3	74.3	79.9	87.7	76.5	81.7	90.1	65.7	76.0	92.1	71.4	77.6
TDRG [35]	(576, 576)	86.0	87.0	74.7	80.4	87.5	77.9	82.4	90.7	65.6	76.2	91.9	68.0	78.1
CCD-R101 [20]	(576, 576)	85.3	88.3	73.1	80.2	88.8	76.3	82.1	91.0	65.2	76.0	92.3	67.3	77.9
Ours(SADCL)	(448, 576)	86.8	86.4	77.0	81.1	87.7	79.1	83.2	90.0	67.4	75.7	92.0	68.7	78.7

Comparisons with state-of-the-art methods on the MS-COCO dataset.* indicates the reproduced results of our implementation. All metrics are in %.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [25]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
ResNet-101* [12]	99.0	98.4	97.5	96.0	81.4	97.3	97.3	97.1	79.6	96.0	88.1	97.5	98.5	95.8	98.8	85.9	97.2	84.6	98.8	92.0	93.8
RNN-Att [26]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
SSGRL⋆ [2]	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.1	84.9	96.5	79.8	98.4	92.8	93.4
ML-GCN [5]	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
P-GCN [6]	99.6	98.6	98.4	98.7	81.5	94.8	97.6	98.2	83.1	96.0	87.1	98.3	98.5	96.3	99.1	87.3	95.5	85.4	98.9	93.6	94.3
ADDGCN⋆ [33]	99.8	99.0	98.4	99.0	86.7	98.1	98.5	98.3	85.8	98.3	88.9	98.8	99.0	97.4	99.2	88.3	98.7	90.7	99.5	97.0	96.0
KGGR⋆ [1]	99.3	98.6	97.9	98.4	86.2	97.0	98.0	99.2	82.6	98.3	87.5	99.0	98.9	97.4	99.1	86.9	98.2	84.1	99.0	95.0	95.0
DSDL [36]	99.8	98.7	98.4	97.9	81.9	95.4	97.6	98.3	83.3	95.0	88.6	98.0	97.9	95.8	99.0	86.6	95.9	86.4	98.6	94.4	94.4
GM-MLIC [29]	99.4	98.7	98.5	97.6	86.3	97.1	98.0	99.4	82.5	98.1	87.7	99.2	98.9	97.5	99.3	87.0	98.3	86.5	99.1	94.9	94.7
TDRG [35]	99.9	98.9	98.4	98.7	81.9	95.8	97.8	98.0	85.2	95.6	89.5	98.8	98.6	97.1	99.1	86.2	97.7	87.2	99.1	95.3	95.0
MulCon⋆ [8]	99.8	98.3	99.3	98.6	83.3	98.4	98.0	98.3	85.8	98.3	90.5	99.3	98.9	96.6	98.8	86.3	99.8	87.3	99.8	96.1	95.6
CPCL [30]	99.6	98.6	98.5	98.8	81.9	95.1	97.8	98.2	83.0	95.5	85.5	98.4	98.5	97.0	99.0	86.6	97.0	84.9	99.1	94.3	94.4
SST [3]	99.8	98.6	98.9	85.5	94.7	97.9	98.6	83.0	96.8	85.7	98.8	98.8	98.9	95.7	99.1	85.4	96.2	84.3	99.1	95.0	94.5
Ours(SADCL)⋆	100.0	99.0	99.5	99.1	88.9	98.8	98.7	99.6	84.2	98.4	90.1	99.4	99.6	99.0	99.3	90.2	99.6	88.9	99.8	95.3	96.4

Comparisons with state-of-the-art methods on the VOC 2007 dataset. * indicates the reproduced results of our implementation, and ⋆ indicates the model pre-trained on MS-COCO. All of the inputs are 448×448 resolution except SSGRL(576), ADDGCN(512) and KGGR(576).

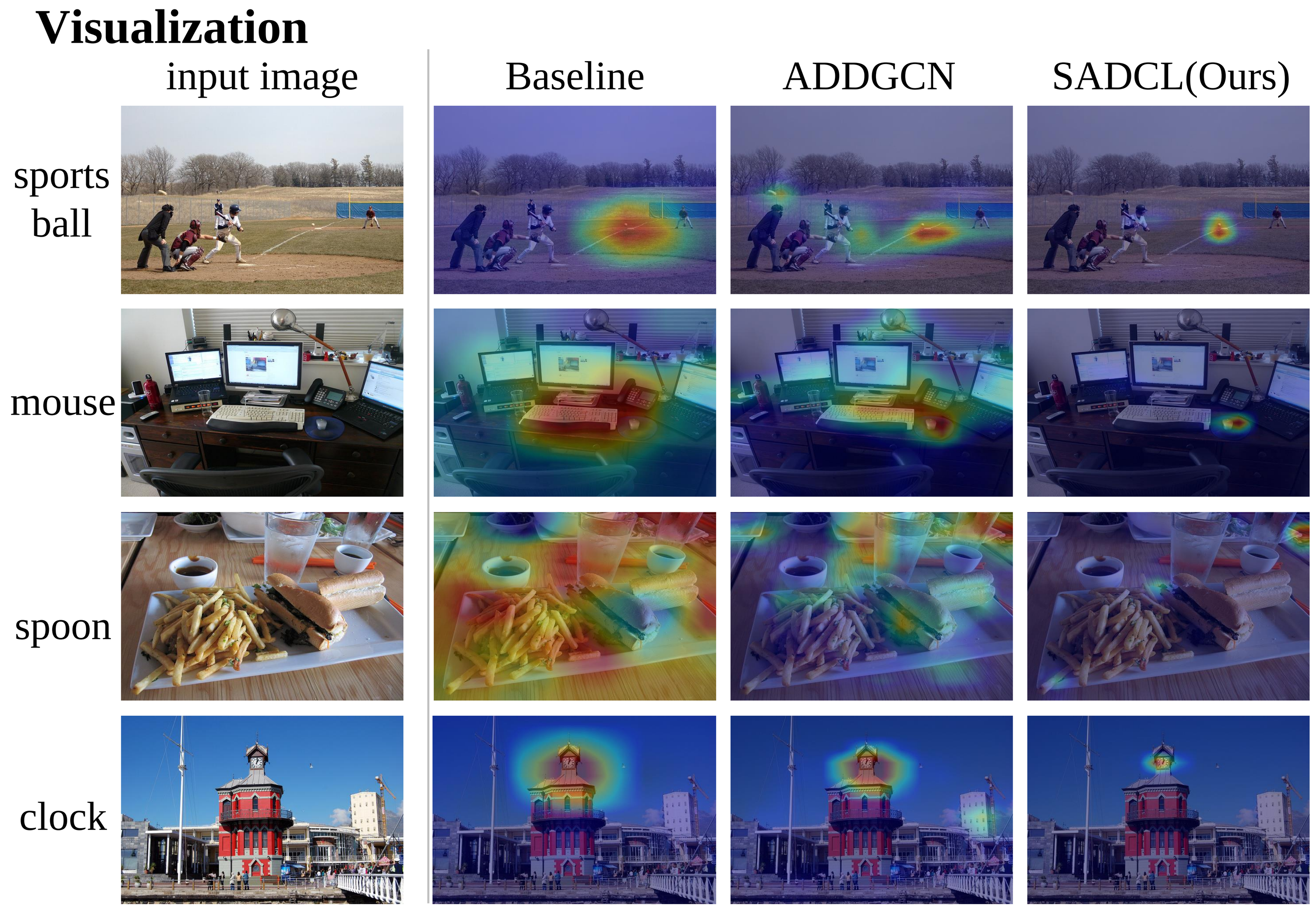
Module Baseline	SARL*	SARL	SSCL	PSCL	MS-COCO mAP	MS-COCO OF1	MS-COCO CF1	NUS-WIDE mAP	NUS-WIDE OF1	NUS-WIDE CF1
✓					81.5	79.7	76.0	62.5	73.8	59.2
✓	✓				84.7	81.5	78.7	65.0	74.2	61.0
✓		✓			85.0	81.7	79.3	65.3	74.8	61.7
✓		✓	✓		85.4	81.7	79.6	65.7	74.9	62.7
✓		✓		✓	85.4	82.1	79.8	65.6	75.0	62.4
✓		✓	✓	✓	**85.6**	**82.1**	**79.8**	**65.9**	**75.0**	**63.0**

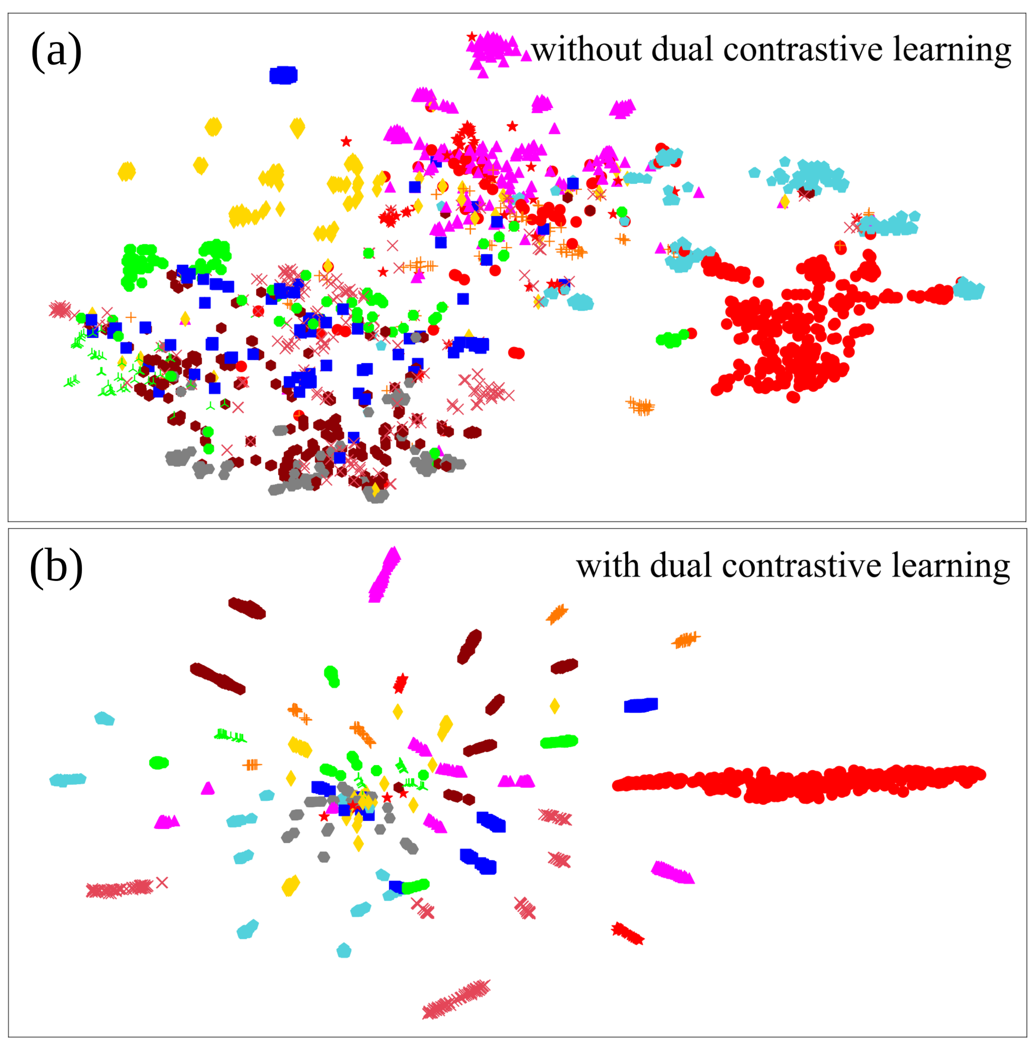
Ablation study of different components in MS-COCO and NUS-WIDE Datasets. Baseline mean ResNet101 network with average pooling and fc layer. SARL* denotes SARL after removing Transformer-encoder. All metrics reflect all predicted scores instead of taking the top-3 highest prediction scores.

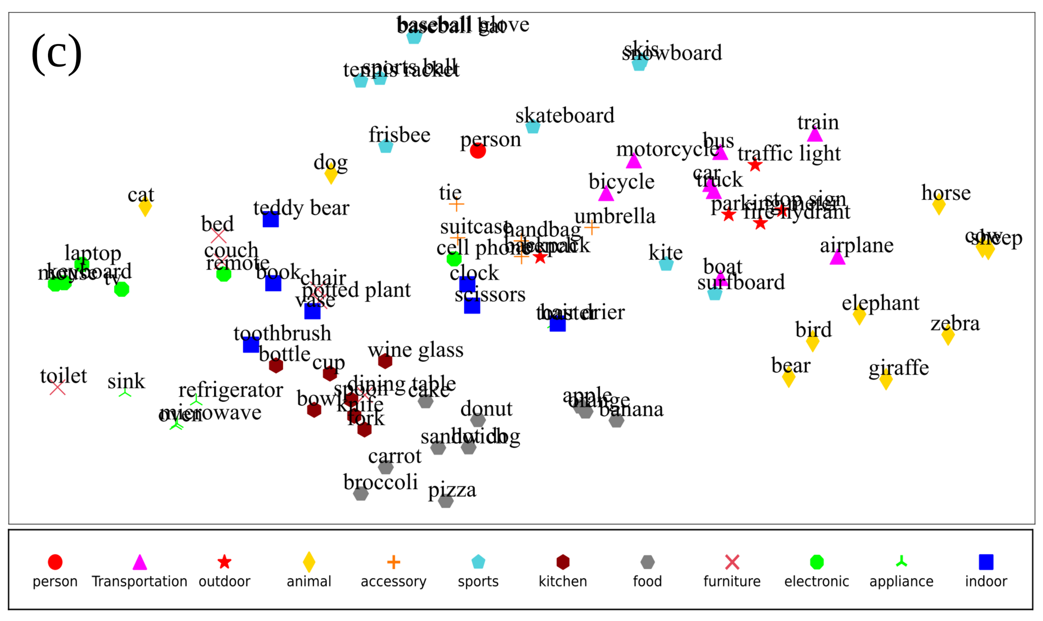
Methods	mAP	All CF1	All OF1	Top 3 CF1	Top 3 OF1
CNN-RNN [25]	-	-	-	34.7	55.2
ResNet101* [12]	62.5	59.2	73.8	54.6	69.4
CADM [4]	62.8	60.7	74.1	56.3	70.6
ADDGCN* [33]	63.3	60.3	73.5	56.5	69.3
P-GCN [6]	62.8	60.4	73.4	57.0	69.1
TDRG* [35]	63.5	60.0	73.8	56.1	69.5
SST [3]	63.5	59.6	73.2	55.9	68.8
MulCon [8]	63.9	61.8	74.8	-	-
CCD-R101* [20]	64.2	61.8	74.6	56.7	70.0
Ours(SADCL)	**65.9**	**63.0**	**75.0**	**57.8**	**70.6**

Comparisons with state-of-the-art methods on the NUSWIDE dataset. * indicates the reproduced results of our implementation. The input images are resized to 448×448 resolution in both the training and testing phases.

Visualization







Figures (a) and (b) visualize 2000 randomly sampled label-level visual representations from the MS-COCO test dataset. Figure (c) visualizes the learned category prototypes on the MS-COCO dataset. Different colors and shapes represent different superclasses.