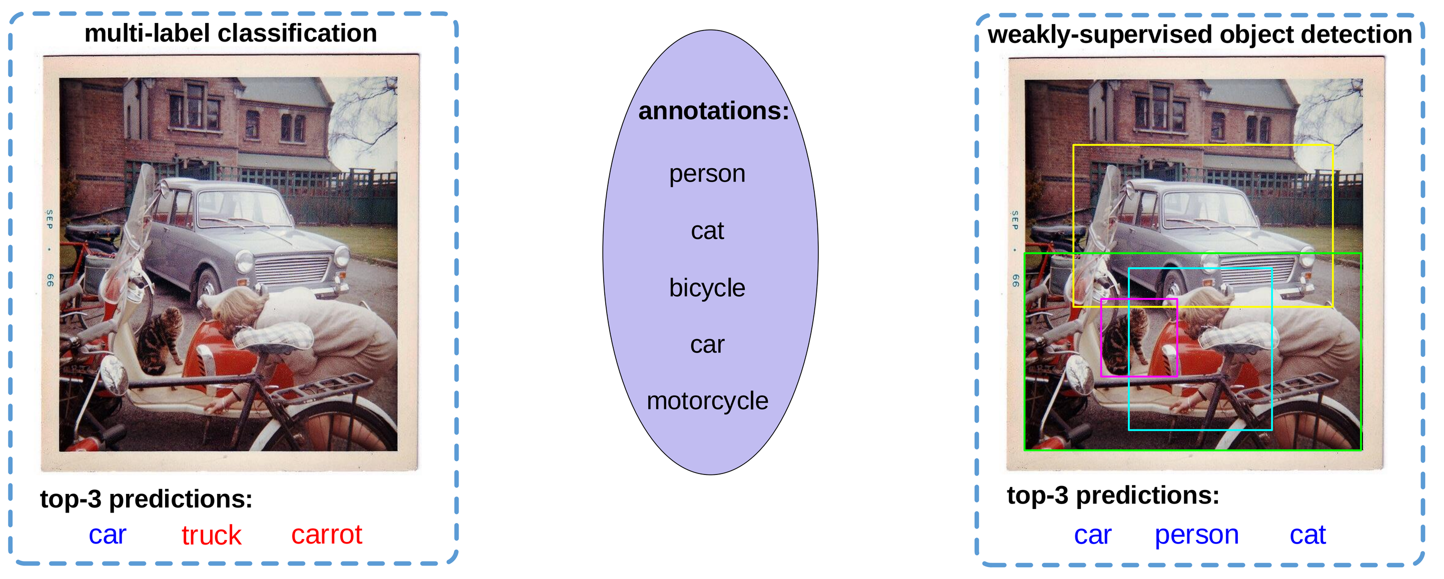
ACM MULTIMEDIA 2018
22 - 26 October 2018, Seoul, Korea

Multi-Label Image Classification via Knowledge Distillation from Weakly-Supervised Detection

Yongcheng Liu[1,2], Lu Sheng[3], Jing Shao[4], Junjie Yan[4], Shiming Xiang[1,2], Chunhong Pan[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong [4]SenseTime Research

Motivation

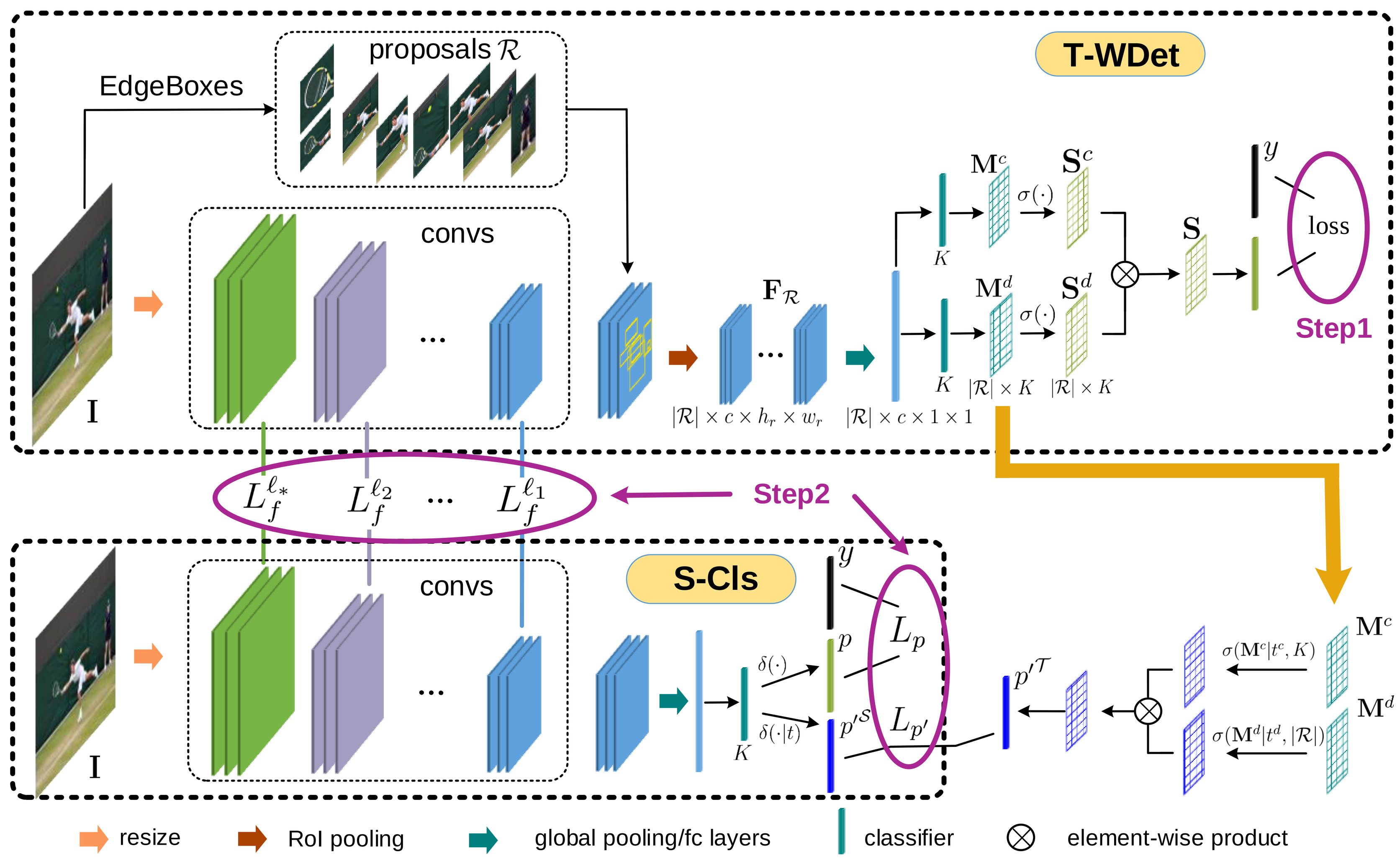


- The **Multi-Label Image Classification (MLIC)** model can not work well due to *poor localization* for multiple semantic instances.
- The detections by **Weakly-Supervised Detection (WSD)** model tend to locate the *semantic regions* which are *informative for classifying* the target object, although they may not preserve object boundaries well.
- The localizations of WSD could provide *object-relevant informative regions*, the image-level predictions of WSD could capture the *latent class dependencies*, both can facilitate the MLIC task.

Contribution

- A novel deep MLIC framework equipped with *cross-task knowledge distillation*, *i.e.*, distilling the unique knowledge from WSD into MLIC.
- The first work that applies *knowledge distillation between two different tasks*, *i.e.*, weakly-supervised detection and multi-label image classification.
- Extensive experiments on two challenging large-scale datasets (MS-COCO and NUS-WIDE) demonstrate the effectiveness of the proposed framework.

Overall Framework



A novel deep framework to boost MLIC by *distilling the unique knowledge* from WSD into classification with only image-level annotations. The WSD is taken as the **teacher** (**T-WDet**) while the MLIC is the **student** (**S-Cls**).

Step 1: Weakly-Supervised Detection

We first develop a WSD model with image-level annotations (WSDDN in this paper).

Step 2: Cross-Task Knowledge Distillation (WSD is frozen)

Stage 1: Feature-level transfer. Distilling the object-relevant features from RoIs.

Minimize $\sum_\ell L_f^\ell(\mathbf{w}^{\mathcal{S}}_{\text{conv}})$

only update convs' params

$$\begin{cases} L_f(\mathbf{w}^{\mathcal{S}}_{\text{conv}}) = \frac{1}{2N}\sum_n \frac{1}{|\mathcal{R}'_n|}\|\mathbf{F}^{\mathcal{T}}_{\mathcal{R}'_n} \ominus \mathbf{F}^{\mathcal{S}}_{\mathcal{R}'_n}\|_2^2 \\ \mathbf{F}^{\mathcal{T}}_{\mathcal{R}'_n} = C_{R\in\mathcal{R}'_n}\left[s'_R \odot \phi_{\text{RoI}}(\mathbf{F}^{\mathcal{T}}_{\text{conv}}; R)\right], \\ \mathbf{F}^{\mathcal{S}}_{\mathcal{R}'_n} = C_{R\in\mathcal{R}'_n}\left[s'_R \odot \phi_{\text{RoI}}(\Psi(\mathbf{F}^{\mathcal{S}}_{\text{conv}})|\mathbf{w}^{\mathcal{S}}_{\text{conv}}; R)\right] \end{cases}$$

Stage 2: Prediction-level transfer. Distilling the class dependencies from image-level predictions of WSD.

Minimize $L_p(\mathbf{w}^{\mathcal{S}}) + \lambda L_{p'}(\mathbf{w}^{\mathcal{S}})$

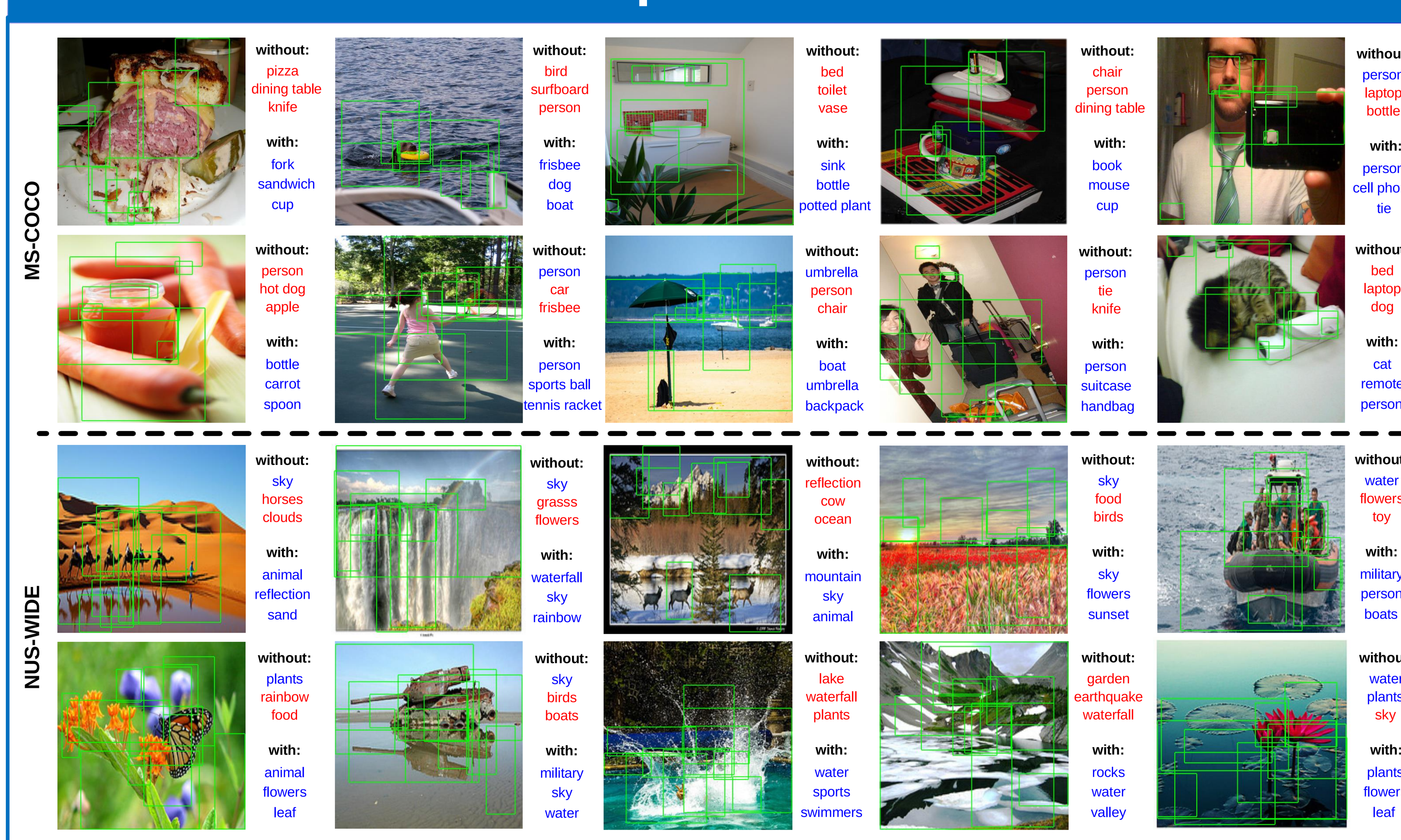
update all params

$$\begin{cases} L_p(\mathbf{w}^{\mathcal{S}}) = -\frac{1}{N}\sum_n[y\log p + (1-y)\log(1-p)] \\ L_{p'}(\mathbf{w}^{\mathcal{S}}) = \frac{1}{2N}\sum_n\|p'^{\mathcal{T}} - p'^{\mathcal{S}}(\mathbf{w}^{\mathcal{S}})\|_2^2 \end{cases}$$

Advantages:

- After cross-task distillation, the MLIC model can be improved significantly.
- It is efficient as the WSD model can be safely discarded in the test phase.

Experiment



MS-COCO: The image in 1st column of the 1st row. After distillation, even the *highly occluded objects* like "fork" and "cup" can be well recognized.

NUS-WIDE: The image in 2nd column of the 1st row. After distillation, *motion and event concepts* like "waterfall" and "rainbow" are recognized.

Table 1: Quantitative comparison (%) on MS-COCO.

Method	All			Top-3	
	mAP	F1-C	F1-O	F1-C	F1-O
CNN-RNN [32]	-	-	-	60.4	67.8
CNN-LSEP [19]	-	62.9	68.3	-	-
CNN-SREL-RNN [21]	-	63.4	<u>72.5</u>	-	-
RMAM(512+10crop) [33]	72.2	-	-	<u>66.5</u>	<u>71.3</u>
RARLF(512+10crop) [5]	-	-	-	65.6	70.5
MIML-FCN-BB [39]	66.2	-	-	-	-
MCG-CNN-LSTM [43]	64.4	-	-	58.1	61.3
RLSD [43]	68.2	-	-	62.0	66.5
Ours-S-Cls (w/o)	70.9	63.6	67.0	60.7	66.7
Distillation [12]	71.3	64.7	69.3	61.5	67.6
FitNets [23]	<u>72.5</u>	<u>65.2</u>	70.9	62.3	68.3
Attention transfer [42]	71.4	64.6	69.8	61.6	67.8
Ours-S-Cls (w/)	**74.6**	**69.2**	**74.0**	**66.8**	**72.7**

Table 2: Quantitative comparison (%) on NUS-WIDE.

Method	All			Top-3	
	mAP	F1-C	F1-O	F1-C	F1-O
CNN-RNN [32]	-	-	-	34.7	55.2
Tag-Neighbors [15]	52.8	-	-	45.2	62.5
CNN-LSEP [19]	-	52.9	70.8	-	-
CNN-SREL-RNN [21]	-	52.8	<u>71.0</u>	-	-
MCG-CNN-LSTM [43]	52.4	-	-	46.1	59.9
RLSD [43]	54.1	-	-	46.9	60.3
KCCA [30]	52.2	-	-	-	-
Ours-S-Cls (w/o)	55.6	52.0	67.2	47.5	64.8
Distillation [12]	57.2	54.3	69.5	50.3	67.5
FitNets [23]	57.4	54.9	70.4	51.4	68.6
Attention transfer[42]	<u>57.6</u>	<u>55.2</u>	70.3	<u>51.7</u>	<u>68.8</u>
Ours-S-Cls (w/)	**60.1**	**58.7**	**73.7**	**53.8**	**71.1**

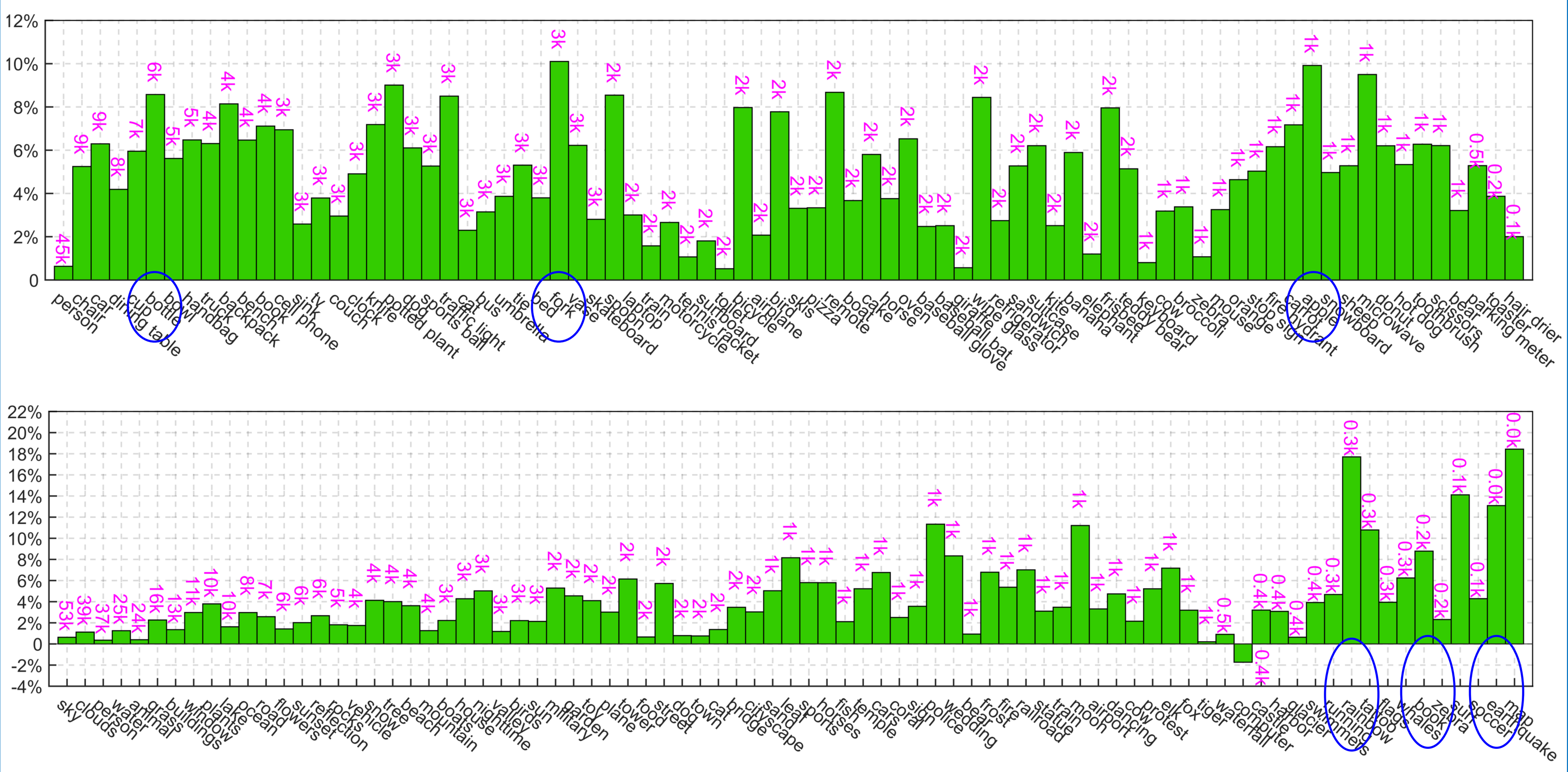


Figure 4: The improvements over each class on MS-COCO (upper) and NUS-WIDE (lower) after knowledge distillation. "*k" indicates the number (divided by 1000) of images including this class.

- The improvements are also considerable even when the classes are very *imbalanced*.
- The framework are robust to the *object's size* and the *label's type*.
  - ✓ On MS-COCO, small objects like "bottle", "fork", "apple" and so on, which may be difficult for the classification model to pay attention, are also improved a lot.
  - ✓ On NUS-WIDE, scenes (e.g., "rainbow"), events (e.g., "earthquake") and objects (e.g., "book") are all improved considerably.

Ablation Study

Table 3: Overall ablation study.

Dataset	mAP		
	S-Cls (w/o)	T-WDet	S-Cls (w/)
MS-COCO	70.9	78.6	74.6
NUS-WIDE	55.6	58.2	60.1

Table 4: Component-wise ablation study.

Method	mAP
Baseline (Sigmoid-Logistic)	70.9
+Distillation [12]	71.3
+Class-aware distillation	72.1
+NMS proposals transfer+Class-aware transfer	73.8
+RoI-aware transfer+Class-aware transfer	74.6

Table3:

- ✓ the MLIC model not only obtains *global information* learned from annotations,
- ✓ but also perceives the local *object-relevant regions* as *complementary cues* distilled from the WSD model,
- ✓ thus it could surpass the teacher (WSD) on NUS-WIDE.

Table 5: Region proposals from EdgeBoxes and Faster-RCNN.

Method	mAP
Baseline (Sigmoid-Logistic)	70.9
T-WDet (EdgeBoxes [47])	78.6
S-Cls	74.6
T-WDet (Faster RCNN [22])	81.1
S-Cls	76.3

Table5:

- ✓ EdgeBoxes: unsupervised
- ✓ Faster-RCNN: supervised
- ✓ 74.6 vs 76.3, the gap is not obvious

Information

Paper **Project Page** **Code**

Contact information: yongcheng.liu@nlpr.ia.ac.cn (Yongcheng Liu)
shaojing@sensetime.com (Jing Shao)