

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

Yongcheng Liu<sup>1,2</sup>, Lu Sheng<sup>3</sup>, Jing Shao<sup>4,\*</sup>, Junjie Yan<sup>4</sup>, Shiming Xiang<sup>1,2</sup>, Chunhong Pan<sup>1</sup>

<sup>1</sup> National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

<sup>2</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences

<sup>3</sup> CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong <sup>4</sup> SenseTime Research

{yongcheng.liu, smxiang, chpan}@nlpr.ia.ac.cn, lsheng@ee.cuhk.edu.hk, {shaojing\*, yanjunjie}@sensetime.com

<https://yongchengliu.github.io/MLIC-KD-WSD/>

## ABSTRACT

Multi-label image classification is a fundamental but challenging task towards general visual understanding. Existing methods found the region-level cues (e.g., features from RoIs) can facilitate multi-label classification. Nevertheless, such methods usually require laborious object-level annotations (i.e., object labels and bounding boxes) for effective learning of the object-level visual features. In this paper, we propose a novel and efficient deep framework to boost multi-label classification by distilling knowledge from weakly-supervised detection task without bounding box annotations. Specifically, given the image-level annotations, (1) we first develop a weakly-supervised detection (WSD) model, and then (2) construct an end-to-end multi-label image classification framework augmented by a knowledge distillation module that guides the classification model by the WSD model according to the class-level predictions for the whole image and the object-level visual features for object RoIs. The WSD model is the *teacher* model and the classification model is the *student* model. After this cross-task knowledge distillation, the performance of the classification model is significantly improved and the efficiency is maintained since the WSD model can be safely discarded in the test phase. Extensive experiments on two large-scale datasets (MS-COCO and NUS-WIDE) show that our framework achieves superior performances over the state-of-the-art methods on both performance and efficiency.

## KEYWORDS

Multi-Label Image Classification, Weakly-Supervised Detection, Knowledge Distillation

## 1 INTRODUCTION

Multi-label image classification (MLIC) [7, 29] is one of the pivotal and long-lasting problems in computer vision and multimedia. This task starts from the observation that real-world images always contain diverse semantic contents that need multiple visual concepts to classify. Except for the challenges shared with single-label image

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

© 2018 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

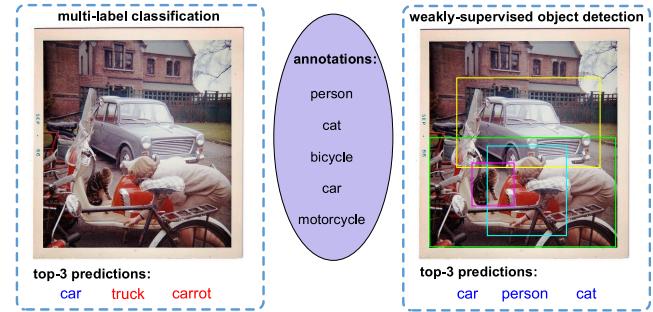


Figure 1: The illustration of multi-label image classification (MLIC) and weakly-supervised detection (WSD). We show top-3 predictions, in which correct predictions are shown in blue and incorrect predictions in red. The MLIC model might not predict well due to poor localization for semantic instances. Although the detection results of WSD may not preserve object boundaries well, they tend to locate the semantic regions which are informative for classifying the target object, such that the predictions can still be improved.

classification (e.g., inter-class similarity and intra-class variation), MLIC is more difficult because predicting the presence of multiple classes usually needs a more thorough understanding of the input image (e.g., associating classes with semantic regions and capturing the semantic dependencies of classes).

Contemporary methods may simply finetune the multi-label classification networks pre-trained on the single-label classification datasets (e.g., ImageNet [24]). However, the classifiers trained for global image representations may not generalize well to the images in which objects from multiple classes are distributed in different locations, scales and occlusions. To mitigate this problem, the task of MLIC can be decomposed into multiple independent binary classification tasks, in which one classifier only focuses on one object label. In this way, though very efficient, the semantic dependencies among multiple classes, which is especially important for MLIC [32], are ignored (e.g., “cat” is more likely to be misclassified into the category of “dog” than falsely associated to “car”). Therefore, some prior works [4, 32, 38] tried to fix this drawback by explicitly capturing the class dependencies with a RNN or LSTM structure appended after CNN-based models. However, they usually suffer from the difficulty in back-propagating stable gradients [21].

Recently, some object localization techniques [34, 43, 44] are introduced into the MLIC task by simplifying the multi-label classification problem into multi-object detection task. The resulting pipeline usually involves two steps. The hypothesis regions are first proposed using low-level image cues [31]. Then a neural network is

trained to predict class scores on these proposals, and these predictions are aggregated to achieve MLIC task. Even though satisfactory performance can be achieved with sufficiently accurate region proposal algorithms, these methods always have to bear redundant computational cost in the test phase. Thus they are usually not practical for large-scale applications.

To solve above issues, an effective and efficient multi-class image classification model needs to simultaneously hold three important advantages: (1) locating semantic regions for object-level feature extraction; (2) capturing semantic dependencies among multiple classes; (3) fewer additional computation and annotation budgets for the practical issue. Following this intuition, weakly-supervised detection (WSD) [2] may be a feasible solution. It could achieve the detection goal of locating each semantic instance with a specific class using only image-level annotations. Figure 1 shows the task illustrations for MLIC and WSD frameworks. The MLIC model might not predict well due to the lack of object-level feature extraction and localization for semantic instances. Although the results detected by WSD may not preserve object boundaries well, they tend to locate the semantic regions which are informative for classifying the target object, such that the predictions can still be improved. Therefore, the localization results of WSD could provide object-relevant semantic regions while its image-level predictions could naturally capture the class dependencies. These unique advantages are very useful for the MLIC task. The only problem is the huge computational complexity in the WSD pipelines. Is it possible to combine the advantages in WSD with the high efficiency of simple classification network? Knowledge distillation [12], a technique that distills knowledge from a large *teacher* model into a much smaller *student* model, may provide a good solution to guide the classification model to inherently contain object-level localization ability and mutual class dependencies.

In this paper, we propose a novel and efficient deep framework to boost MLIC by distilling the unique knowledge from WSD into classification with only image-level annotations. The overall architecture of our framework is illustrated in Figure 2. Specifically, our framework works with two steps: (1) we first develop a WSD model with image-level annotations; (2) then we construct an end-to-end knowledge distillation framework by propagating the class-level holistic predictions and the object-level features from RoIs in the WSD model to the MLIC model, where the WSD model is taken as the *teacher* model (called T-WDet) and the classification model is the *student* model (called S-ClS). The distillation of object-level features from RoIs focuses on perceiving localizations of semantic regions detected by the WSD model while the distillation of class-level holistic predictions aims at capturing class dependencies predicted by the WSD model. After this distillation, the classification model could be significantly improved and no longer need the WSD model, thus resulting in high efficiency in test phase.

The main contributions of this work are highlighted as follows:

- A novel and efficient deep multi-label image classification framework equipped by knowledge distillation is proposed, which distills the unique knowledge from a weakly-supervised detection model into the classification model such that the latter is improved significantly with high efficiency.
- To our best knowledge, it is the first work that applies knowledge distillation between two different tasks, i.e., weakly-supervised detection and multi-label image classification.
- Extensive experiments are conducted on two large-scale public datasets (MS-COCO and NUS-WIDE), and the results show that our framework achieves superior performances over the state-of-the-art methods on both performance and efficiency.

## 2 RELATED WORK

**Multi-Label Image Classification (MLIC).** The progress of MLIC [9, 28, 34, 45] has been greatly made with deep convolutional neural network [11, 16, 25]. Some works [13, 15, 18, 19, 21, 27, 32, 37, 38, 41, 44] embed label dependencies with the deep model to improve the accuracy of MLIC. CNN-RNN [32] utilizes RNN combined with CNN to learn a joint image-label embedding for capturing label dependencies. [21] proposes a regularised embedding layer as the interface between the CNN and RNN to mitigate the difficulty of model training in [32]. [18, 27, 35] learn graph structure to model the label dependencies. These methods always require pre-defined label relations. Some other works ensemble multiple deep models with different input scales [5, 33] while suffering high complexity.

Recently, various methods [4, 5, 33, 39, 40, 43, 45] have been proposed to locate semantic regions for learning deep attentional representations. For example, MIML-FCN+ [39] uses bounding boxes from Faster-RCNN [22] to locate the objects in an image for multi-instance learning. Spatial regularization network [45] generates class-related attention maps to capture spatial dependencies. [33] employs a LSTM sub-network to predict labeling scores on the regions located by a spatial transformer layer.

All the aforementioned methods either require pre-defined label relations or object-level annotations, and they usually add model complexity by the extra modules, both of which result in poor practicality.

**Weakly-supervised detection (WSD).** Recently, many researches on WSD [2, 8, 17, 26, 46] have been conducted. Dual-network [8] is proposed to optimize proposal generation and instance selection in a joint framework. [17] introduces the domain adaptation techniques for the WSD task. WSDDN [2] modifies ImageNet pre-trained VGG [25] to operate at image regions, performing simultaneously region selection and classification. In this work, we extend the WSDDN to develop a WSD model into our framework.

**Knowledge Distillation.** Hinton et al. [12] use a softened version of the output of a large *teacher* network to teach information to a small *student* network. FitNets [23] employs not only the output but also intermediate layer values of the teacher network to train the student network. Attention transfer [42] forces the student network to be consistent with the teacher network on feature attention maps. These methods focus on the distillation between the same tasks, and they always use the whole feature maps and class-identical soften targets to conduct distillation, which can not locate to semantic regions of the image and are not sensitive to classes. Chen et al. [3] concentrate on distilling between the same tasks of object detection while our proposed distillation is operated between two different tasks, i.e., from WSD to MLIC.

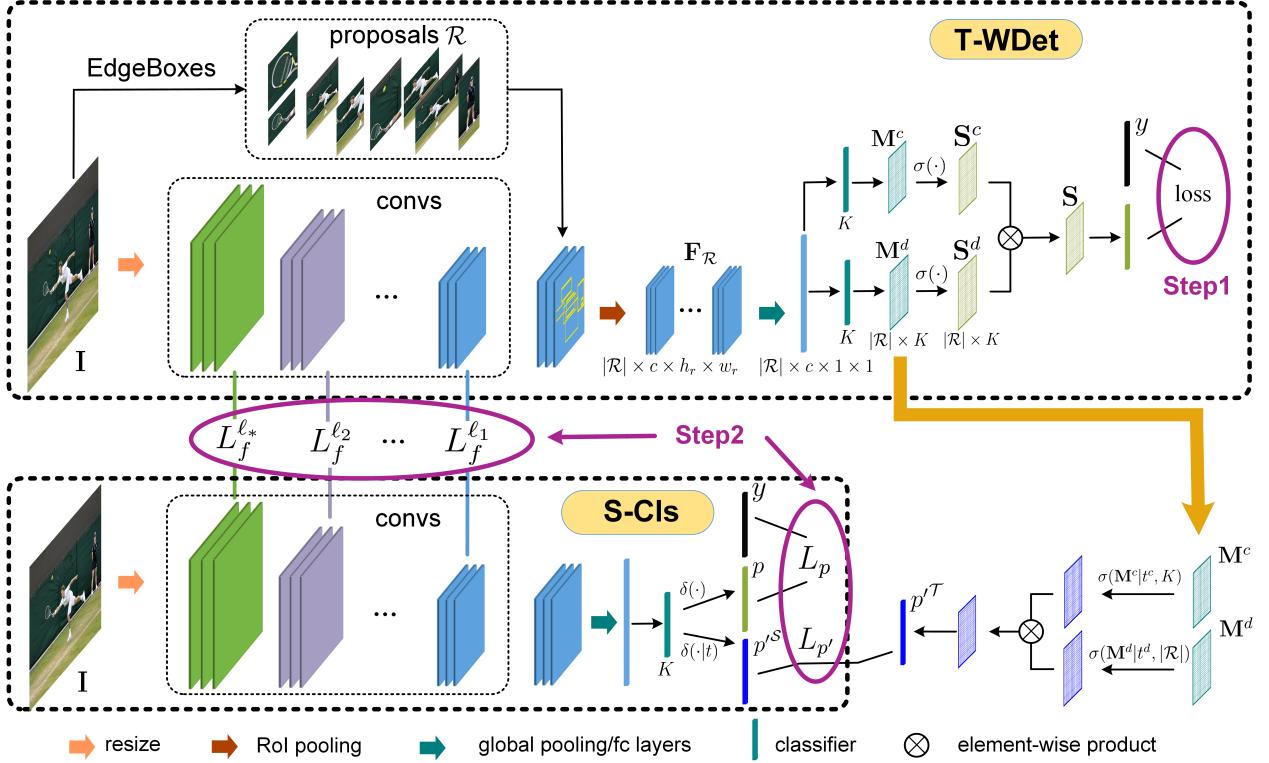


Figure 2: The overall architecture of our framework. The proposed framework works with two steps: (1) we first develop a WSD model as teacher model (called T-WDet) with only image-level annotations  $y$ ; (2) then the knowledge in T-WDet is distilled into the MLIC student model (called S-Cls) via feature-level distillation from RoIs and prediction-level distillation from whole image, where the former is conducted by optimizing the loss  $\sum_\ell L_f^\ell$  while the latter is conducted by optimizing the loss  $L_p$  and  $L_{p'}$ .

### 3 METHODOLOGY

Multi-label image classification aims at obtaining all the semantic classes in an image. Generally, given an image  $I$ , the final prediction  $l_k$  of the  $k$ -th class corresponding to  $I$  is formulated by

$$l_k = \mathbb{I}(p_k(I|\mathbf{w}) > \tau_k), \quad k \in \{1, \dots, K\}, \quad (1)$$

where  $p_k(I|\mathbf{w})$ , estimated by a model with parameters  $\mathbf{w}$ , denotes the posterior probability of image  $I$  including the  $k$ -th class.  $K$  is the number of given labels, and  $\tau_k$  is the confidence threshold for the  $k$ -th class.  $\mathbb{I}(p > \tau)$  is an indicator function, it takes 1 when  $p > \tau$  and 0 otherwise.  $l_k$  is the final label indicator, i.e.,  $l_k = 1$  means the  $k$ -th class is included in the given image and  $l_k = 0$  otherwise.

In this paper, we propose a novel and efficient framework in which the multi-label image classification (MLIC) task is facilitated by weakly-supervised detection task. In the following, we will present the proposed framework in detail, including (1) weakly-supervised detection (WSD) model, (2) Knowledge distillation from WSD to MLIC, and (3) Implementation details.

#### 3.1 Weakly-Supervised Detection (WSD) Model

Although any existing WSD methods can be used in our framework, we choose WSDDN [2] because of its architecture accessibility. Using VGG16 [25] pre-trained on ImageNet [24] as backbone network, WSDDN operates on image regions which are outputted by Edge-Boxes (EB) [47]. In this work, we extend WSDDN to support any popular networks. The architecture of extended WSDDN (called

T-WDet) is illustrated at the upper part of Figure 2. First, EB algorithm is used to get a lot of proposals  $\mathcal{R}$  from the input image  $I$ . These proposals are inputted to the ROI pooling [22] module to get ROI-localized features. Note that we replace SPP pooling [10] by ROI pooling, because the latter keeps the spatial information. Formally, let  $\mathbf{F}_{\text{conv}} \in \mathbb{R}^{c \times h \times w}$  denote the last convolutional feature maps of the backbone network,  $R$  denote a proposal in  $\mathcal{R}$ , and  $s_R$  denote the prior score of  $R$  outputted by EB algorithm,  $\mathbf{F}_R \in \mathbb{R}^{|\mathcal{R}| \times c \times h_r \times w_r}$ , the obtained ROI features of all the proposals, can be described as

$$\begin{aligned} \mathbf{F}_R &= s_R \odot \phi_{\text{ROI}}(\mathbf{F}_{\text{conv}}; R), \\ \mathbf{F}_R &= \mathbb{C}_{R \in \mathcal{R}}(\mathbf{F}_R), \end{aligned} \quad (2)$$

where  $\phi_{\text{ROI}}(\cdot)$  is the operation of ROI pooling, “ $\odot$ ” is the operation of multiplying each element of  $\phi_{\text{ROI}}(\cdot)$  by score  $s_R$ , and  $\mathbb{C}(\cdot)$  is the concatenation operation, which concatenates the features of  $|\mathcal{R}|$  proposals along the fourth dimension.

Then, global pooling or several fully connected (fc) layers are adopted to further transform the ROI feature maps  $\mathbf{F}_R$  into feature vectors. The subsequent network is split into two branches. Both of them pass through a fully connected layer, where the output is consistent with the given classes  $K$ , to get a logit matrix  $\mathbf{M} \in \mathbb{R}^{|\mathcal{R}| \times K}$ . One branch aims at classification while the other at detection. The classification is achieved by a softmax operation along the first dimension  $K$ , and the detection along the second dimension  $|\mathcal{R}|$ .

Finally, the element-wise product operation is adopted to fuse the softmax score matrix  $\mathbf{S}^c, \mathbf{S}^d \in \mathbb{R}^{|\mathcal{R}| \times K}$  of the two branches.

The fused score matrix  $\mathbf{S}$  is summed along the second dimension  $|\mathcal{R}|$  to get final class prediction  $p \in \mathbb{R}^K$ , which is compatible with the image-level annotations  $y \in \mathbb{R}^K$ . The final detection results for each class are obtained by processing each column of  $\mathbf{S}$  with non-maximum suppression (NMS). T-WDet is also trained in an end-to-end manner. More details can be referred in [2].

### 3.2 Knowledge Distillation from WSD to MLIC

In this paper, we argue that there is unique knowledge beyond classification contained in the task of WSD, which could facilitate MLIC. Specifically, on one hand, the detection results of WSD provide localization of semantic regions, which is a powerful cue for classification model to further understand the image. On the other hand, the image-level prediction confidences of WSD naturally capture semantic dependencies among classes, which could be a strong reference for MLIC from the perspective of detection.

As stated in Section 3.1, we first use a T-WDet model to achieve the goal of detection. The problem locates at how to transfer the unique knowledge from T-WDet model into the classification model. A reasonable solution is knowledge distillation [12], which can distill the knowledge in a large *teacher* model for improving a small *student* model. Inspired by this idea, we propose a dedicated distillation framework to distill knowledge from a WSD *teacher* model (T-WDet) for boosting a MLIC *student* model (S-Clss). This distillation framework works with two stages. The first stage focuses on the feature-level knowledge transfer while the second stage on the prediction-level knowledge transfer. Both of the two stages are included in “step 2” in Figure 2.

**Feature-level knowledge transfer.** We propose a ROI-aware distillation approach which explicitly distills the localization knowledge from WSD to MLIC at feature level. Specifically, we sum the fused score matrix  $\mathbf{S}$  in the well-trained T-WDet model along the first dimension  $K$  to get a confidence vector  $s' \in \mathbb{R}^{|\mathcal{R}|}$ . This vector implies a confidence distribution of all the proposals’ objectness, i.e., region proposal score, which is a reliable localization importance indicator for MLIC. Since the proposals outputted by EB algorithm are highly overlapped, we take NMS operation for them using the obtained confidences  $s'$ . Then, with these well-chosen proposals, the knowledge from T-WDet model is distilled into S-Clss model by minimizing the  $\ell_2$  loss of ROI pooled features on selected convolutional layers as

$$L_f(\mathbf{w}_{\text{conv}}^S) = \frac{1}{2N} \sum_n \frac{1}{|\mathcal{R}'_n|} \|\mathbf{F}_{\mathcal{R}'_n}^{\mathcal{T}} \ominus \mathbf{F}_{\mathcal{R}'_n}^S\|_2^2, \quad (3)$$

where  $\mathcal{R}'_n$  denotes the remaining proposals after performing NMS to  $\mathcal{R}_n$  for image  $I_n$  and  $N$  is the number of training images. “ $\ominus$ ” is the element-wise subtraction operation.  $\mathbf{F}_{\mathcal{R}'_n}^{\mathcal{T}}$  and  $\mathbf{F}_{\mathcal{R}'_n}^S$  denote the ROI pooled features from T-WDet model and S-Clss model, respectively. They can be described as

$$\begin{aligned} \mathbf{F}_{\mathcal{R}'_n}^{\mathcal{T}} &= \mathbb{C}_{R \in \mathcal{R}'_n} [s'_R \odot \phi_{\text{ROI}}(\mathbf{F}_{\text{conv}}^{\mathcal{T}}; R)], \\ \mathbf{F}_{\mathcal{R}'_n}^S &= \mathbb{C}_{R \in \mathcal{R}'_n} [s'_R \odot \phi_{\text{ROI}}(\Psi(\mathbf{F}_{\text{conv}}^S) | \mathbf{w}_{\text{conv}}^S; R)], \end{aligned} \quad (4)$$

where  $\mathbf{F}_{\text{conv}}^{\mathcal{T}}$  and  $\mathbf{F}_{\text{conv}}^S$  denote the selected convolutional layers in T-WDet model and S-Clss model, respectively.  $\Psi(\mathbf{F}_{\text{conv}}^S)$  is the possibly needed transforming operation, which transforms  $\mathbf{F}_{\text{conv}}^S$  to be compatible with  $\mathbf{F}_{\text{conv}}^{\mathcal{T}}$  in case the number of their channels is

different. In this process, we only update the convolutional parameters  $\mathbf{w}_{\text{conv}}^S$  in S-Clss model and  $s'_R$  plays a role as local importance weighting factor for proposal  $R$ .

Our ROI-aware distillation approach explicitly distills the unique knowledge from the detection results of T-WDet model into S-Clss model, i.e., localization of semantic regions and objectness confidence. It is superior to FitNets [23] and attention transfer [42], because both of them transfer knowledge on whole feature map, which is not sensitive to localization and objectness. Our distillation approach can also be operated on multiple layers, then the loss we minimize becomes  $\sum_\ell L_f^\ell(\mathbf{w}_{\text{conv}}^S)$ .

**Prediction-level knowledge transfer.** The final label prediction  $p$  of T-WDet model is obtained by summing the score matrix  $\mathbf{S}$  along the second dimension  $|\mathcal{R}|$ . It aggregates the confidence of all the proposals over the given classes, which is a powerful reference for classification. Moreover, we observe that the classification accuracy for different classes between T-WDet model and S-Clss model are very different, thus the prediction-knowledge transfer should be of difference over classes. To discriminatively distill the knowledge from T-WDet model to S-Clss model at prediction level, we propose a class-aware distillation approach. Specifically, after initializing the parameters  $\mathbf{w}^S$  of S-Clss model with  $\mathbf{w}_{\text{conv}}^S$  pre-trained in the first stage, we then simultaneously minimize two different loss functions for S-Clss model in this stage. The first loss function is the  $\ell_2$  loss of the discriminatively softened predictions of T-WDet model and S-Clss model as

$$L_{p'}(\mathbf{w}^S) = \frac{1}{2N} \sum_n \|p'^{\mathcal{T}} - p'^S(\mathbf{w}^S)\|_2^2, \quad (5)$$

where  $p'^{\mathcal{T}}$  is the softened predictions of T-WDet model, which is calculated by

$$p'^{\mathcal{T}} = \sum_{i=1}^{|\mathcal{R}|} [\sigma(\mathbf{M}^c | t^c, K) \otimes \sigma(\mathbf{M}^d | t^d, |\mathcal{R}|)], \quad \mathbf{M} \in \mathbb{R}^{|\mathcal{R}| \times K}. \quad (6)$$

Here, “ $\otimes$ ” is the element-wise product operation.  $\sigma(\mathbf{M}^c | t^c, K)$  and  $\sigma(\mathbf{M}^d | t^d, |\mathcal{R}|)$  are the softened softmax operation along the first dimension  $K$  (classification branch) and the second dimension  $\mathcal{R}$  (detection branch) on the logit matrix  $\mathbf{M}$ , respectively. They can be defined as

$$\begin{aligned} [\sigma(\mathbf{M}^c | t^c, K)]_{ij} &= \frac{e^{m_{ij}^c / t_k^c}}{\sum_{k=1}^K e^{m_{ik}^c / t_k^c}}, \quad \forall i \in \{1, \dots, |\mathcal{R}|\}, \\ [\sigma(\mathbf{M}^d | t^d, |\mathcal{R}|)]_{ij} &= \frac{e^{m_{ij}^d / t_r^d}}{\sum_{r=1}^{|\mathcal{R}|} e^{m_{rj}^d / t_r^d}}, \quad \forall j \in \{1, \dots, K\}, \end{aligned} \quad (7)$$

where  $t_k^c$  and  $t_r^d$  are the softmax temperature of  $k$ -th class and the softmax temperature of  $r$ -th proposal, respectively.

$p'^S(\mathbf{w}^S)$  is the softened sigmoid predictions of S-Clss model, which is calculated by

$$p'^S(\mathbf{w}^S) = \delta(m_k | t) = 1 / (1 + e^{-m_k / t_k}), \quad (8)$$

where  $m_k$  and  $t_k$  are the logit and the sigmoid temperature of  $k$ -th class, respectively. We decompose multi-label classification task as multiple binary classification tasks, and we use sigmoid operation to get the final output.

---

**Algorithm 1** Training and Test of S-Cls model

---

**TRAINING**

**Input:** image data and label data ( $I^N, y^N$ ).

**Output:** parameters  $w$  of S-Cls model.

**Initialize:**  $w, \lambda$  and training hyper-parameters.

**Stage 1:** Feature-Level Knowledge Transfer.

1: **Repeat:**

2: compute  $L_f(w^S_{\text{conv}})$  by Eq. (3), Eq. (4).

3: update  $w^S_{\text{conv}}$  by gradient back-propagation.

4: **Until:**  $L_f(w^S_{\text{conv}})$  converges.

**Stage 2:** Prediction-Level Knowledge Transfer.

5: **Repeat:**

6: compute  $L_p(w^S) + \lambda L_{p'}(w^S)$  by Eq. (5), Eq. (10), Eq. (11).

7: update  $w^S$  by gradient back-propagation.

8: **Until:**  $L_p(w^S)$  converges.

**Return:**  $w^S$

**TEST**

**Input:** image data  $I^N$ .

**Output:** prediction  $l^N$ .

**Initialize:** parameters  $w$  of S-Cls model, confidence threshold  $\tau$ .

9: **for**  $n = 1$  to  $N$  **do**

10: forward pass S-Cls model to get  $p(I^n | w)$ .

11: compute  $l^n$  by Eq. (1).

12: **end for**

**Return:**  $l^N$

---

Note that all the temperatures are different, and they are learnable in the training phase. This is more reasonable for our task than the class-identical and fixed temperature used in [12]. Moreover, it also cuts down the laborious costs for tuning the artificial temperatures by this learnable way. Formally, let  $m$  denote the input data,  $t$  denote the temperature and  $\hat{m}$  denote the output data:  $\hat{m}_i = m_i / t_i$ , then the back-propagation and chain rule are used to compute derivatives w.r.t  $m$  and  $t$  as

$$\frac{\partial L_{p'}}{\partial m_i} = \sum_{t_i} \frac{\partial L_{p'}}{\partial \hat{m}_i} \frac{1}{t_i}, \quad \frac{\partial L_{p'}}{\partial t_i} = \sum_{m_i} \frac{\partial L_{p'}}{\partial \hat{m}_i} \left(-\frac{m_i}{t_i^2}\right). \quad (9)$$

The second loss function is the cross entropy with hard label (ground truth)  $y$  as

$$L_p(w^S) = -\frac{1}{N} \sum_n [y \log p + (1 - y) \log(1 - p)], \quad (10)$$

where  $p$  is the normal sigmoid prediction of S-Cls model.

In the class-aware distillation stage, we update all the parameters  $w^S$  of S-Cls model. For one thing, the S-Cls model fits the given hard labels by working as multiple binary classification tasks. For another, it also acquires the knowledge of detection, i.e., semantic dependencies of classes distilled from well-trained T-WDet model.

### 3.3 Implementation Details

**Training.** In the training phase, we first train a T-WDet model as stated in Section 3.1. Then, we froze all the parameters of well-trained T-WDet model, and train the S-Cls model using the proposed ROI-aware and class-aware distillation framework. This is operated

with two stages. **Stage 1:** We train S-Cls model to update convolutional parameters  $w^S_{\text{conv}}$  by optimizing the loss in Eq. (3). **Stage 2:** We update the whole network by optimizing the weighted losses in Eq. (10) and Eq. (5) as

$$L_p(w^S) + \lambda L_{p'}(w^S), \quad (11)$$

where  $\lambda$  is the weighted factor.

**Test.** In the test phase, the S-Cls model works without T-WDet model. It is compact as the same as standard classification model, i.e., no any extra computational cost. The normal sigmoid outputs  $p$  is taken as its final predictions. The pseudo-code of training and test of S-Cls model can be referred in Algorithm 1.

To convincingly demonstrate the proposed framework, we use VGG16 pre-trained on ImageNet as backbone network for both T-WDet and S-Cls models. VGG16 is the most popular network used in the literature of MLIC, thus a fair comparison can be made. The ROI pooling size is set to  $7 \times 7$  for the two networks. The image size input to S-Cls model is always  $224 \times 224$ .

**T-WDet model.** For training with mini-batch, we take top 500 proposals of EB algorithm, which are sorted by the prior scores of EB. Moreover, we recycle high-score proposals if the candidate number is less than 500. Except mirror flip, we train T-WDet model also using the popular techniques applied in detection area, i.e., training with multiple square scales of 480, 576, 672, 736 and 832. We use the input scale of 672 in the distillation process.

**Knowledge distillation.** In feature-level knowledge transfer, the NMS threshold is set to 0.4 to clean highly overlapped proposals. Moreover, we take top 100 proposals after NMS for training with mini-batch, and again, we recycle them if the candidate number is not enough. In this stage, we only use the conv5\_3 layer for knowledge transfer. The transforming operation is set to  $\Psi(F) = F$  due to the equal number of channels between two networks. In prediction-level knowledge transfer, the convolutional layers of S-Cls model are initialized with  $w^S_{\text{conv}}$  trained in last stage, fully connected layers are initialized with ImageNet pre-trained parameters, and other layers are initialized with Xavier algorithm [36]. The value of all the temperatures  $t$  are initialized with 1. The weighted factor  $\lambda$  of two losses is set to 1.

Our framework is implemented using Caffe [14]. The stochastic gradient descend (SGD) algorithm is employed for the network training, with a batchsize of 32, momentum of 0.9 and weight decay of 0.0005. For feature-level transfer, the learning rate is fixed at  $10^{-5}$  and the training continues for about 100 epochs. For prediction-level transfer, the initial learning rate is set to  $10^{-4}$ , and decreased to  $1/10$  when validation loss gets saturated.

## 4 EXPERIMENT

### 4.1 Datasets

The proposed framework is evaluated on two large-scale datasets with fairly different types of labels: MS-COCO [20] with 80 object labels and NUS-WIDE [6] with 81 concept labels.

**MS-COCO.** It contains 122,218 images of 80 object labels, with about 2.9 labels per image. The objects are of high diversity, and they are of severe occlusions. We follow the official split of 82,081 images for training and 40,137 validation images for testing.

**NUS-WIDE.** This dataset contains 269,648 images and 5018 tags from Flickr. There are a total of 1000 tags after removing noisy

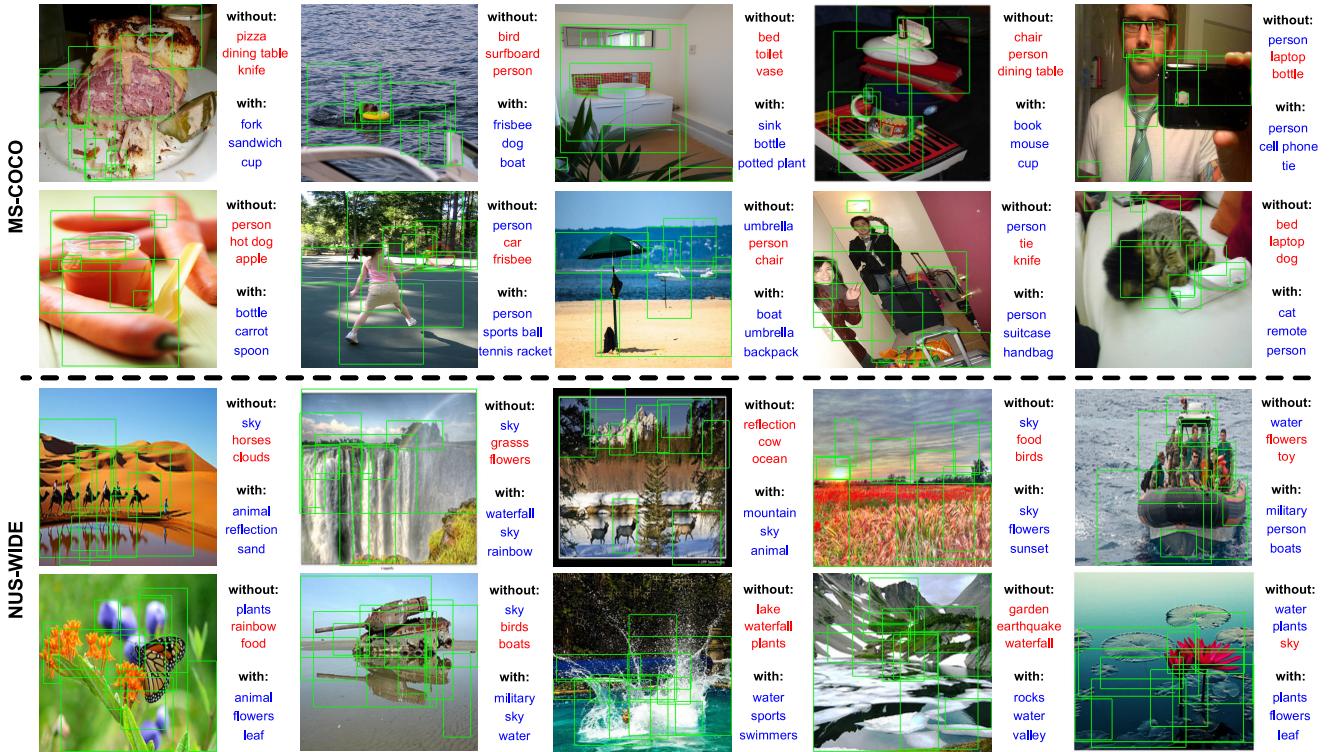


Figure 3: Example results on two datasets. The green bounding boxes in images are the top-10 proposals detected by T-WDet model, which is sorted by objectness confidences  $s'$  in Eq. 4. The text on the right of images are the top-3 classification results of S-Cls model “without” and “with” knowledge distillation using our framework, where correct predictions are shown in blue and incorrect predictions in red.

Table 1: Quantitative comparison (%) on MS-COCO. “w/” and “w/o” indicate “with” and “without” knowledge distillation by the proposed framework, respectively. The values in bold are the best while the values underlined are the second best.

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/)	<b>74.6</b>	<b>69.2</b>	<b>74.0</b>	<b>66.8</b>	<b>72.7</b>

and rare tags. These images are further manually annotated into 81 concepts with 2.4 concepts per image on average. The concepts are quite diverse, including event (e.g., running), scene/location (e.g., airport), object (e.g., animal). We follow the split used in [9, 21], i.e., 150,000 images for training and 59,347 for testing after removing the images without any labels.

Note that both of the two datasets are imbalanced over classes, and the imbalance on NUS-WIDE is even worse.

Table 2: Quantitative comparison (%) on NUS-WIDE. “w/” and “w/o” indicate “with” and “without” knowledge distillation by the proposed framework, respectively. The values in bold are the best while the values underlined are the second best.

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]	57.6	<u>55.2</u>	70.3	<u>51.7</u>	<u>68.8</u>
Ours-S-Cls (w/)	<b>60.1</b>	<b>58.7</b>	<b>73.7</b>	<b>53.8</b>	<b>71.1</b>

## 4.2 Evaluation Metrics and Compared Methods

**Evaluation Metrics.** We employ three overall metrics for comparison: macro/micro F1 (“F1-C”/“F1-O”) and mean average precision (mAP). Macro F1 is evaluated by averaging per-class F1, while micro F1 is evaluated on the results of all the images over all classes. For computing F1, we tune a class-independent confidence threshold, i.e., if the confidence is greater than this threshold, the prediction is taken as positive. We also report top-3 F1 sorted by the prediction confidences. mAP is the mean precision over classes.

Generally, mAP is of more reference, because it directly measures ranking quality and does not require choosing the final predictions.

**Compared Methods.** We compare our framework against the following stage-of-the-art deep learning methods: **CNN-RNN** [32] and **CNN-SREL-RNN** [21] employ CNN combined with RNN for classification; **CNN-LSEP** [19] estimates the optimal confidence thresholds for each class; **RMAM** [33] and **RARLF** [5] locate to image regions for classification, they use very large input size ( $512 \times 512$ ) and multi-scale and multi-crop tricks in the test phase; **MIML-FCN-BB** [39] uses outputs from Faster RCNN [22] with bounding box annotations for classification; **RLSD** [43] employs RNN to capture dependencies at localized regions for classification; **Tag-Neighbors** [15] uses CNN to blend information from the image and its neighbors; **MCG-CNN-LSTM** [43] employs LSTM to capture dependencies at proposals of MCG [1]. Note that for fair comparison, we only report the results of methods based on VGG16 network and results without using extra label information (e.g., detailed metadata in NUS-WIDE dataset) or ensemble testing (e.g., fusion of multi-scale and multi-crop test).

We also compare with three advanced distillation methods by implementing them following their paper: (1) Distillation [12]: the  $t$  in T-WDet model and S-Cls model are tuned at 1, 5 and 2, 5 for MS-COCO and NUS-WIDE, respectively. (2) FitNets [23]: we choose the middle layer conv3\_3 as hint layer, then training with the setting in (1). (3) Attention transfer [42]: as suggested in the paper, we choose conv3\_3, conv4\_3 and conv5\_3 as transfer layers, and the transfer is combined with (1).

### 4.3 Experimental Results

**MS-COCO.** Experimental results on this dataset are summarised in Table 1. With ImageNet pre-training, the simple S-Cls model can also perform well, and it outperforms the state-of-the-arts after knowledge distillation using our framework. Compared with RMAM [33] and RARLF [5], which rely on a large input size and multi-crop test, our S-Cls model still performs better with small input ( $224 \times 224$ ) and single-forward test. Moreover, our framework with only image-level annotations also outperforms those methods like MIML-FCN-BB, which require bounding box annotations. The framework also shows superior performance over other advanced distillation methods even though all of them get decent results. Some example results are shown at the upper part of Figure 3. As can be seen, although the proposals of T-WDet hold poor preserve of object boundaries, they locate at the semantic regions that are very informative for classification, thus the classification results are greatly improved. Taking the 1<sup>st</sup> column of the 1<sup>st</sup> row as an example, we can see the objects in the image are highly overlapped, resulting in a poor classification result, the top-3 predictions are all wrong. However, after distillation with detected semantic regions, even the occluded objects “fork” and “cup” can be well recognised.

**NUS-WIDE.** Quantitative results on this dataset are summarised in Table 2. The simple S-Cls model again performs well with a backbone network pre-trained on ImageNet, and it also outperforms the state-of-the-arts after knowledge distillation with our framework. Meanwhile, compared with other architectures that add extra modules to the backbone network, e.g., CNN-RNN [32], CNN-SREL-RNN [21] and RLSD [43], our framework performs better with higher efficiency at the same time. Moreover, the proposed

framework also outperforms other advanced distillation methods. We also show some example results at the lower part of Figure 3. As it shows, the T-WDet model can still locate semantic regions even with the concept label, and the classification results are again improved by distilling these informative regions into the classification model. Taking the 2<sup>nd</sup> column of the 1<sup>st</sup> row as an example, the classification model only recognises correctly with one concept “sky” by a global perception of this image in top-3 predictions. However, other concepts like “waterfall” and “rainbow” are recognised after the distillation with our framework. This demonstrates the effectiveness and robustness of our framework simultaneously.

The improvements over each class on two datasets are shown in Figure 4. As it shows, on one hand, the improvements on MS-COCO are relatively even to the classes, while NUS-WIDE focuses on the classes in which the number of images is fewer. This demonstrates the effectiveness of our framework to mitigate the problem of data imbalance, since the NUS-WIDE dataset is very imbalanced (the number of images on “sky” is 53k while quite a few concepts only hold hundreds of images). On the other hand, the improvements on MS-COCO focus on small objects like “bottle”, “fork”, “apple” and so on, which may be difficult for the classification model to pay attention. This indicates the importance of semantic regions where T-WDet model is distilled into S-Cls model, which is shown in Figure 3. Moreover, on NUS-WIDE, the improvements focus on scenes (e.g., “rainbow”), events (e.g., “earthquake”) and objects (e.g., “book”), which demonstrates the robustness of our framework to the types of labels.

### 4.4 Ablation Study

**Overall ablation study.** The results of overall ablation study on two datasets are summarised in Table 3. The T-WDet model achieves very good performance on MS-COCO while slightly better performance on NUS-WIDE. The main reason is that the clean object labels on MS-COCO are quite suitable for detection task while the noisy concept labels are not. Moreover, the S-Cls model is improved on both datasets after knowledge distillation by our framework, which verifies its effectiveness.

**Component-wise ablation study.** We also perform component-wise ablation study on MS-COCO to carefully evaluate the contribution of the critical components of our framework. The baseline is the VGG16-based S-Cls model trained with sigmoid logistic loss as Eq. 10. The results are summarised in Table 4. It improves a little when directly applying the distillation methods proposed by [12]. After adding our class-aware distillation approach to baseline, the performance is improved much more (from 71.3 to 72.1). This demonstrates that our discriminative knowledge distillation at prediction level is superior than class-identical distillation [12].

We then add feature-level knowledge distillation followed with class-aware distillation in the way of two-stage training. The performance is improved considerably (from 72.3 to 73.8) when we take NMS operation to all the proposals based on their objectness confidences, and it improves again when weighting the localized features with these confidences. This demonstrates that the objectness confidence obtained from T-WDet model is a reliable indicator of semantic of the proposal.

We also analyse our framework by the experiment using supervised detection results. Specifically, we replace the EB proposals

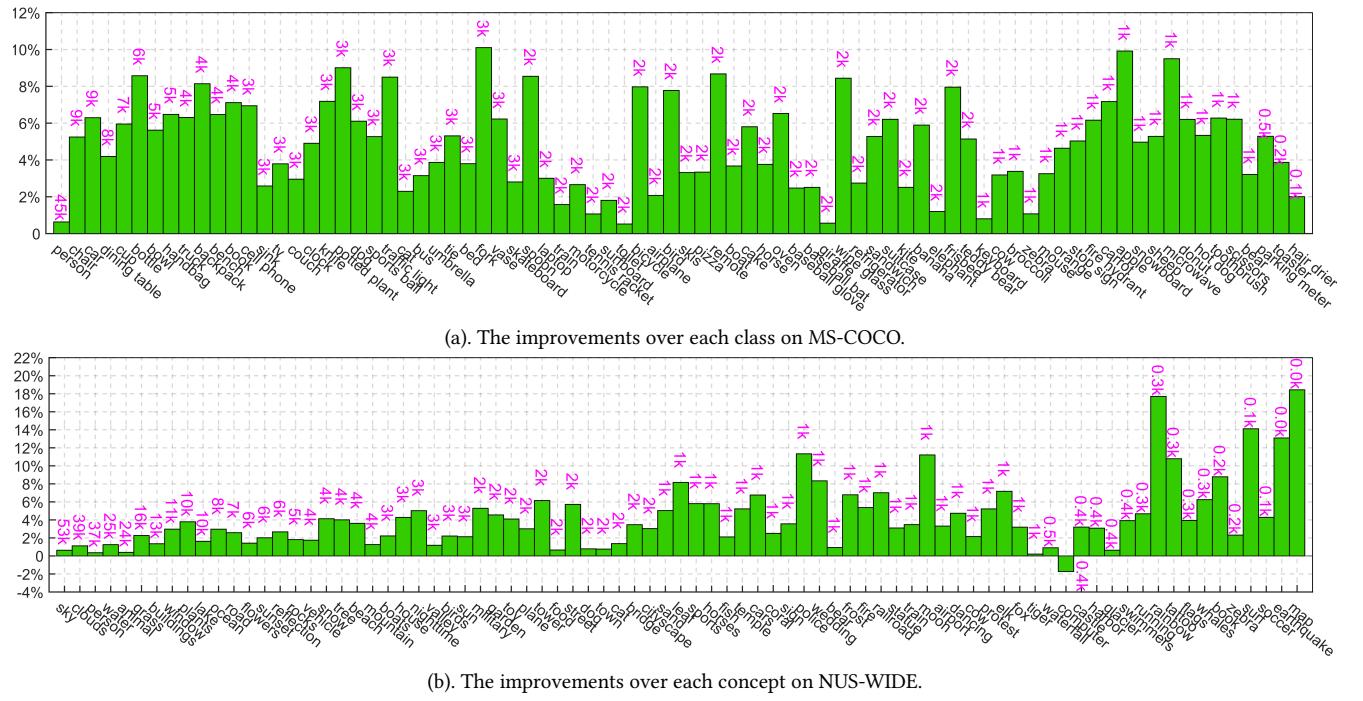


Figure 4: The improvements of S-Cls model over each class/concept on two datasets after knowledge distillation with our framework. “**k**” indicates the number (divided by 1000) of images including this class/concept. The classes/concepts in horizontal axis are sorted by the number “**k**” from large to small.

Table 3: Overall ablation study on two datasets (%). “w/” and “w/o” indicate “with” and “without” knowledge distillation by the proposed framework, respectively.

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

input to T-WDet model by the detection results from Faster RCNN [22]. All the hyper-parameters are set to the same as the source code of [22], which results in 100 proposals for each image. The results are summarised in Table 5. As can be seen, the classification performance of T-WDet is improved from 78.6 to 81.1 when using the supervised detection results. After distillation with our framework, S-Cls model is improved to 76.3 compared with EB proposals to 74.6, where the gap is not obvious. This further demonstrates the effectiveness of our proposed framework.

## 5 CONCLUSION

In this paper, a novel and efficient deep framework for multi-label image classification has been proposed. It boosts classification by distilling the unique knowledge from weakly-supervised detection

**Table 5: The comparison of region proposals from EdgeBoxes [47] and Faster-RCNN [22] (%).**

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

(WSD) into classification with only image-level annotations. The proposed framework works with two steps. A WSD model is first developed, then an end-to-end knowledge distillation framework is constructed via feature-level distillation from RoIs and class-level distillation from predictions, where the WSD model is the *teacher* model and the classification model is the *student* model. The feature-level distillation from RoIs learns to perceiving semantic regions detected by the WSD model while class-level distillation aims at capturing class dependencies in the predictions of the WSD model. Thanks to this effective distillation, the classification model could be significantly improved without the WSD model in the test phase, thus resulting in high efficiency. Extensive experiments on two large-scale public datasets (MS-COCO and NUS-WIDE) show that the proposed framework outperforms the state-of-the-arts on both performance and efficiency. In the future, we will explore the complementary cues that could facilitate weakly-supervised detection to further boost the multi-label classification task.

## ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under grant 91646207.

## REFERENCES

- [1] Pablo Arbelaez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marques, and Jitendra Malik. 2014. Multiscale Combinatorial Grouping. In *IEEE Conference on Computer Vision and Pattern Recognition*. 328–335.
- [2] Hakan Bilen and Andrea Vedaldi. 2016. Weakly Supervised Deep Detection Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2846–2854.
- [3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. 2017. Learning Efficient Object Detection Models with Knowledge Distillation. In *Advances in Neural Information Processing Systems*. 742–751.
- [4] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. 2018. Order-Free RNN with Visual Attention for Multi-Label Classification. In *AAAI Conference on Artificial Intelligence*. 1–8.
- [5] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. 2018. Recurrent Attentional Reinforcement Learning for Multi-label Image Recognition. In *AAAI Conference on Artificial Intelligence*. 1–8.
- [6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2014. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *ACM International Conference on Image and Video Retrieval*. 1–9.
- [7] Radhika Devkar and Sankirti Shiravale. 2017. A Survey on Multi-label Classification for Images. *International Journal of Computer Applications*. 162, 8 (2017), 39–42.
- [8] Xuanyi Dong, Deyu Meng, Victor S. Sheng, Fan Ma, and Yi Yang. 2017. A Dual-Network Progressive Approach to Weakly Supervised Object Detection. In *ACM International Conference on Multimedia*. 279–287.
- [9] Yunchao Gong, Yangqing Jia, Thomas K. Leung, Alexander Toshev, and Sergey Ioffe. 2014. Deep Convolutional Ranking for Multilabel Image Annotation. In *International Conference on Learning Representations*. 1–9.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 37, 9 (2015), 1904–1916.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- [12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*. (2015).
- [13] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. 2016. Learning Structured Inference Neural Networks with Label Relations. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2960–2968.
- [14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *ACM International Conference on Multimedia*. 675–678.
- [15] Justin Johnson, Lamberto Ballan, and Li Fei-Fei. 2015. Love Thy Neighbors: Image Annotation by Exploiting Image Metadata. In *IEEE International Conference on Computer Vision*. 4624–4632.
- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. 1106–1114.
- [17] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. 2017. Weakly Supervised Object Localization with Progressive Domain Adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3512–3520.
- [18] Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. 2016. Conditional Graphical Lasso for Multi-label Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2977–2986.
- [19] Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving Pairwise Ranking for Multi-label Image ClassificationImproving Pairwise Ranking for Multi-label Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3617–3625.
- [20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*. 740–755.
- [21] Feng Liu, Tao Xiang, Timothy M. Hospedales, Wankou Yang, and Changyin Sun. 2017. Semantic Regularisation for Recurrent Image Annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2872–2880.
- [22] Shaqiq Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*. 91–99.
- [23] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representations*. 1–13.
- [24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*. 115, 3 (2015), 211–252.
- [25] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*. 1–14.
- [26] Krishna Kumar Singh and Yong Jae Lee. 2017. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. In *IEEE International Conference on Computer Vision*. 3524–3533.
- [27] Mingkui Tan, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, and Junbin Gao. 2015. Learning graph structure for multi-label image classification via clique generation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4100–4109.
- [28] Peng Tang, Xinggang Wang, Zilong Huang, Xiang Bai, and Wenyu Liu. 2017. Deep Patch Learning for Weakly Supervised Object Classification and Discovery. *Pattern Recognition*. 71 (2017), 446–459.
- [29] Grigoris Tsoumakas and Ioannis Katakis. 2009. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*. 3, 3 (2009), 1–13.
- [30] Tiberio Uricchio, Lamberto Ballan, Lorenzo Seidenari, and Alberto Del Bimbo. 2017. Automatic Image Annotation via Label Transfer in the Semantic Space. *Pattern Recognition*. 71 (2017), 144–157.
- [31] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders. 2011. Segmentation as Selective Search for Object Recognition. In *IEEE International Conference on Computer Vision*. 1879–1886.
- [32] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2285–2294.
- [33] Zhouxiao Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. 2017. Multi-label Image Recognition by Recurrently Discovering Attentional Regions. In *IEEE International Conference on Computer Vision*. 464–472.
- [34] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. 2016. CNN: Single-label to Multi-label. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 38, 9 (2016), 1901–1907.
- [35] Jian Wu, Anqian Guo, Victor S. Sheng, Pengpeng Zhao, Zhiming Cui, and Hua Li. 2017. Adaptive Low-Rank Multi-Label Active Learning for Image Classification. In *ACM International Conference on Multimedia*. 1336–1344.
- [36] Glorot Xavier and Y. Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. *Journal of Machine Learning Research*. PP, 9 (2010), 249–256.
- [37] Pengtao Xie, Ruslan Salakhutdinov, Luntian Mou, and Eric P. Xing. 2017. Deep Determinantal Point Process for Large-Scale Multi-Label Classification. In *IEEE International Conference on Computer Vision*. 473–482.
- [38] Hao Yang, Joey Tianyi Zhou, and Jianfei Cai. 2016. Improving Multi-label Learning with Missing Labels by Structured Semantic Correlations. In *European Conference on Computer Vision*. 835–851.
- [39] Hao Yang, Joey Tianyi Zhou, and Jianfei Cai. 2017. MIML-FCN+: Multi-instance Multi-label Learning via Fully Convolutional Networks with Privileged Information. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1577–1585.
- [40] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. 2016. Exploit Bounding Box Annotations for Multi-label Object Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 280–288.
- [41] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. 2017. Learning Deep Latent Spaces for Multi-Label Classification. In *AAAI Conference on Artificial Intelligence*. 2838–2844.
- [42] Sergey Zagoruyko and Nikos Komodakis. 2017. Pay More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representations*. 1–13.
- [43] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu. 2018. Multi-label Image Classification with Regional Latent Semantic Dependencies. *IEEE Transactions on Multimedia*. PP, 99 (2018), 1–11.
- [44] Junjie Zhang, Qi Wu, Jian Zhang, Chunhua Shen, and Jianfeng Lu. 2017. Kill Two Birds with One Stone: Weakly-Supervised Neural Network for Image Annotation and Tag Refinement. *arXiv preprint arXiv:1711.06998*. (2017).
- [45] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5513–5522.
- [46] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. 2017. Soft Proposal Networks for Weakly Supervised Object Localization. In *IEEE International Conference on Computer Vision*. 1841–1850.
- [47] C. Lawrence Zitnick and Piotr Dollar. 2014. Edge Boxes: Locating Object Proposals from Edges. In *European Conference on Computer Vision*. 391–405.