# Saliency Regularization for Self-Training with Partial Annotations - *Appendix*

Shouwen Wang[1] Qian Wan[2] Xiang Xiang[1*] Zhigang Zeng[1]
[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology;
Key Laboratory of Image Processing and Intelligent Control, Ministry of Education
[2]Wuhan Research Institute of Posts and Telecommunications
[1]{shouwen_hust, xex, zgzeng}@hust.edu.cn [2]w252086746@gmail.com

## A. Proposition Proofs

**Proposition 1 (Logit shifting).** *A sample with $|y'_c - \hat{p}_c| \to 0$ is easy to classify and $|y'_c - \hat{p}_c| \to 1$ is hard to classify, where $y'_c = 1$ for $y_c = 1$, and $y'_c = 0$ for $y_c = -1$. $s_c$ as an adaptive margin on the logit $a_c$ adjusts $\frac{\partial L_c}{\partial a_c}$ of category c, thus addressing the imbalance of easy and hard samples.*

**Proof.** As shown in Eq. (1), Focal loss or Asymmetric loss adds the exponential weight with respect to $p_c$ as a modulating factor to adjust the gradients of hard and easy samples.

$$\begin{cases} L_+ = -(1-p_c)^{\gamma_+} log(p_c), \ y_c = 1, \\ L_- = -p_c^{\gamma_-} log(1-p_c), \ y_c = -1, \end{cases} \tag{1}$$

where $\gamma_+$ and $\gamma_-$ are the positive and negative focusing parameters, respectively.

While our SR adds an adaptive margin on logit to achieve gradient adjustment. $L_c$ with SR goes for $L_+$ or $L_-$ based on known $y_c$, where $L_+$ and $L_-$ are defined as

$$\begin{cases} L_+ = -log(\hat{p}_c) = -log(\frac{1}{1+e^{-(a_c+\alpha s_c)}}), \ y_c = 1, \\ L_- = -log(1-\hat{p}_c) = -log(\frac{1}{1+e^{a_c+\alpha s_c}}), \ y_c = -1. \end{cases} \tag{2}$$

Eq. (3) and Eq. (4) are the gradients of $L_c$ without and with SR for the logit $a_c$, respectively.

$$\begin{cases} \frac{\partial L_+}{\partial a_c} = -\frac{1}{1+e^{a_c}}, \ y_c = 1, \\ \frac{\partial L_-}{\partial a_c} = \frac{1}{1+e^{-a_c}}, \ y_c = -1. \end{cases} \tag{3}$$

$$\begin{cases} \frac{\partial L_+}{\partial a_c} = -\frac{1}{1+e^{a_c+\alpha s_c}}, \ y_c = 1, \\ \frac{\partial L_-}{\partial a_c} = \frac{1}{1+e^{-(a_c+\alpha s_c)}}, \ y_c = -1. \end{cases} \tag{4}$$

*Corresponding author (email to xex@hust.edu.cn).

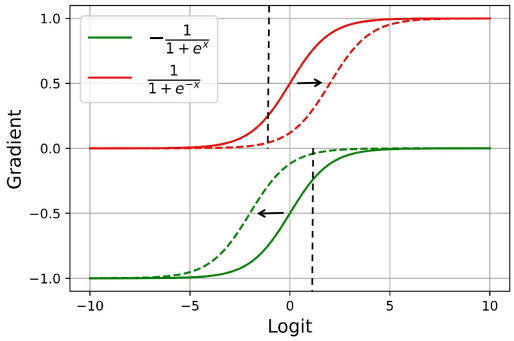


Figure 1. Gradient of BCE. The red and green solid lines represent the gradients of a negative and positive sample in Eq. (3), respectively. The red and green dashed lines are the negative and positive gradients after logit shifting, respectively.

For a easy sample, $|y'_c - \hat{p}_c| \to 0$. When $y_c = 1$, as $\hat{p}_c \to 1$, $s_c > 0$ makes $\frac{\partial L_+}{\partial a_c}$ (negative sign for direction) smaller under the same logit in Eq. (4). When $y_c = -1$, as $\hat{p}_c \to 0$, $s_c < 0$ makes $\frac{\partial L_-}{\partial a_c}$ smaller under the same logit. As a result, easy samples receive less attention. In Fig. 1, with a fixed $s_c$, the gradient changes in different cases are presented for a easy sample.

With a hard sample, $|y'_c - \hat{p}_c| \to 1$. For a hard positive sample, when the sample is misclassified and $\hat{p}_c$ is very small, $s_c < 0$ makes the gradient larger. As $\hat{p}_c \to 0$, $s_c$ gets smaller, but the gradient becomes larger. It is similar for a hard negative sample. Like Focal loss, SR lets the model pay more attention to hard samples. $s_c$ is a adaptive value which relies on different images and features of the same image at different times.

**Proposition 2 (Gradient differentiation).** *For a conventional loss $L_c$ (*i.e.*, BCE, Focal loss or Asymmetric loss) of category c, it propagates the same gradient $\frac{\partial L_c}{\partial a_{c,wh}}$ to each location $(w, h)$ on the c-th CSM. Whereas SR makes*

*the gradient of a location* $(w,h) \in \Omega_c = \{(w,h)|a_{c,wh} \in \mathbf{Topk}(\mathbf{A}_c)\}$ *discriminative with other locations.*

**Proof.** A conventional loss $L_c$ for category $c$ is denoted as

$$L_c = \begin{cases} L_+, \ y_c = 1, \\ L_-, \ y_c = -1. \end{cases} \tag{5}$$

The gradient $\frac{\partial L_c}{\partial a_{c,wh}}$ of a location $(w,h)$ on $\mathbf{A}_c$ is computed as follows:

$$\frac{\partial L_c}{\partial a_{c,wh}} = \frac{\partial L_c}{\partial a_c}\frac{\partial a_c}{\partial a_{c,wh}} = \frac{1}{W \times H}\frac{\partial L_c}{\partial a_c}. \tag{6}$$

The back-propagation gradient of the loss $L_c$ is the same for each location on $\mathbf{A}_c$.

With the addition of SR, the logit changes from $a_c$ to $a_c + \alpha s_c$. $L_c$ is set to BCE, and let $g_c = a_c + \alpha s_c$. The gradient of $L_c$ for $a_{c,wh}$ is denoted as

$$\frac{\partial L_c}{\partial a_{c,wh}} = \frac{\partial L_c}{\partial g_c}\frac{\partial g_c}{\partial a_{c,wh}}. \tag{7}$$

The Eq. (7) is further rewritten as

$$\begin{aligned} \frac{\partial L_+}{\partial a_{c,wh}} &= \frac{\partial L_+}{\partial g_c}\frac{\partial g_c}{\partial a_{c,wh}} = \begin{cases} (\frac{1}{W \times H} + \frac{\alpha}{k})(\hat{p}_c - 1), (w,h) \in \Omega_c, \\ \frac{1}{W \times H}(\hat{p}_c - 1), (w,h) \notin \Omega_c, \end{cases} \\ \frac{\partial L_-}{\partial a_{c,wh}} &= \frac{\partial L_-}{\partial g_c}\frac{\partial g_c}{\partial a_{c,wh}} = \begin{cases} (\frac{1}{W \times H} + \frac{\alpha}{k})\hat{p}_c, (w,h) \in \Omega_c, \\ \frac{1}{W \times H}\hat{p}_c, (w,h) \notin \Omega_c, \end{cases} \end{aligned} \tag{8}$$

where $\Omega_c = \{(w,h)|a_{c,wh} \in \mathbf{Topk}(\mathbf{A}_c)\}$ is a location set. Pixels in the set $\Omega_c$ have larger gradients than other pixels, regardless of the gradient direction. Such optimization makes the features of different locations discriminative. Features of pixels from the set $\Omega_c$ are either emphasized more or suppressed more, which affects the saliency of the object regions for the present labels. From Eq. (8), for a positive sample of category c, the optimization objective is $\frac{\partial L_+}{\partial a_{c,wh}} = 0$, so $\hat{p}_c \to 1$, then $a_c + \alpha s_c \to +\infty \Rightarrow s_c \to +\infty$. For a negative sample of category c, if $\hat{p}_c \to 0$ is obtained, then $s_c \to -\infty$ is the condition ($s_c \to -\infty \Rightarrow a_c + \alpha s_c \to -\infty$). This confirms that adding SR can solve our initial optimization problem.

## B. Dataset Details

Pascal VOC 2007 dataset contains 9963 images with 20 semantic categories. 5011 images are divided into the trainval set and the remaining ones are divided into the test set. MS-COCO dataset covers 80 categories, $123k$ images are divided into $83k$ training images and $40k$ validation images. $108,249$ images and $80,138$ categories are contained in the Visual Genome dataset. We follow the works [3, 8] to select the 200 highest frequency categories to generate a VG-200 subset. For a fair comparison, we utilize the same test set containing $10,000$ images as [3]. The remaining $98,249$ images in the VG-200 are used as training images.

The OpenImages dataset contains 9 million (9M) training images, $41,620$ validation images, and $125,436$ test images. In the main paper, 1.7 million (1.7M) images with $19,693$ categories are downloaded as the training set of OpenImages V3. There are $41,620$ validation images, and $125,436$ test images. Non-annotated images and annotations of non-5000-trainable classes are filtered out, and the remaining images and annotations are used as the final dataset. The dataset has $5,000$ trainable classes and only contains annotations verified by humans. In order to be consistent with the training set of other methods, an additional 1.7 million images are downloaded, for a total of 3.4 million (3.4M) images in the training set. The training set contains less than $0.1\%$ annotated labels, and the positive-negative imbalance is serious in the known labels. It is used to verify the effectiveness of our method further.

## C. Evaluation Metrics

The mean average precision (mAP) over all categories and average of the overall precision, recall, F1-measure (OP, OR, OF1) and per-class precision, recall, F1-measure (CP, CR, CF1) are adopted to evaluate the performance of different methods more comprehensively. These metrics are calculated as follows. For each class, AP is computed as

$$AP_c = \frac{1}{N_c^{gt}}\sum_{k=1}^{N} Precision(k,c) \cdot rel(k,c), \tag{9}$$

where $N_c^{gt}$ is the number of ground-truth images for the $c$-th label, $N$ is the total number of images. $Precision(k,c)$ is the precision for the $c$-th label when retrieving top k predictions, and $rel(k,c)$ is an indicator function that is 1 if the $c$-th label is a positive ground-truth at rank k. The mAP is defined as $mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c$, where $C$ is the number of classes. Other metrics can be computed by

$$OP = \frac{\sum_{c=1}^{C} N_c^{correct}}{\sum_{c=1}^{C} N_c^{predict}}, CP = \frac{1}{C}\sum_{c=1}^{C}\frac{N_c^{correct}}{N_c^{predict}}, \tag{10}$$

$$OR = \frac{\sum_{c=1}^{C} N_c^{correct}}{\sum_{c=1}^{C} N_c^{gt}}, CR = \frac{1}{C}\sum_{c=1}^{C}\frac{N_c^{correct}}{N_c^{gt}}, \tag{11}$$

$$OF1 = \frac{2 \times OP \times OR}{OP + OR}, CF1 = \frac{2 \times CP \times CR}{CP + CR}, \tag{12}$$

where $N_c^{correct}$, $N_c^{predict}$ are the number of correctly predicted images and the number of predicted images for the $c$-th label, respectively.


Figure 2. Qualitative results of complementing unknown labels on several images from the MS-COCO training set. The column of Ground Truth represents main ground-truth positive labels, the column of Known Labels represents the positive labels after randomly dropping the partial labels, and the remaining columns represent the labels complemented by different methods.

## D. Implementation Details

The initial learning rate is set to $0.01$ and divided by $10$ after $15$ epochs with a total of $20$ epochs. In particular, for better convergence, we add $5$ extra epochs for Pascal VOC 2007. The exponential moving average with decay $0.9997$ is utilized for models, like the work [1]. The input image is resized to $448 \times 448$ for training and evaluation. Weak data augmentations include resizing images and flipping images horizontally, and strong data augmentations include Resizing, Flipping, ColorJitter, GaussianBlur, and GridMask [2]. We set hyperparameters $\alpha = 0.5$, $k = 5$, $\tau = 0.6$ on Pascal VOC 2007, MS-COCO, and Visual Genome. Based on the experimental setup and evaluation metric of the work [5], we conduct experiments on the OpenImages V3 benchmark for our method, and also reproduce the results of the work [1] in the main paper. We train the model on $8$ Tesla V100 GPUs with a batch size of $64$, epoch to $15$, and scale of input images to $224 \times 224$. Because of large categories and few known labels in each image of OpenImages V3, pseudo-labels impact the model significantly. Thus, we reduce the weight of the unsupervised loss to $0.1$.

## E. Visualization

As shown in Fig. 2, the labels complemented by different methods are presented. The models are trained by BN-BCE (batch normalization for BCE), Focal loss, and our method on the MS-COCO training set with 10% known labels. Unknown labels are predicted by the models on several training images. Compared to other methods, our method complements more labels. Some images do not have known labels

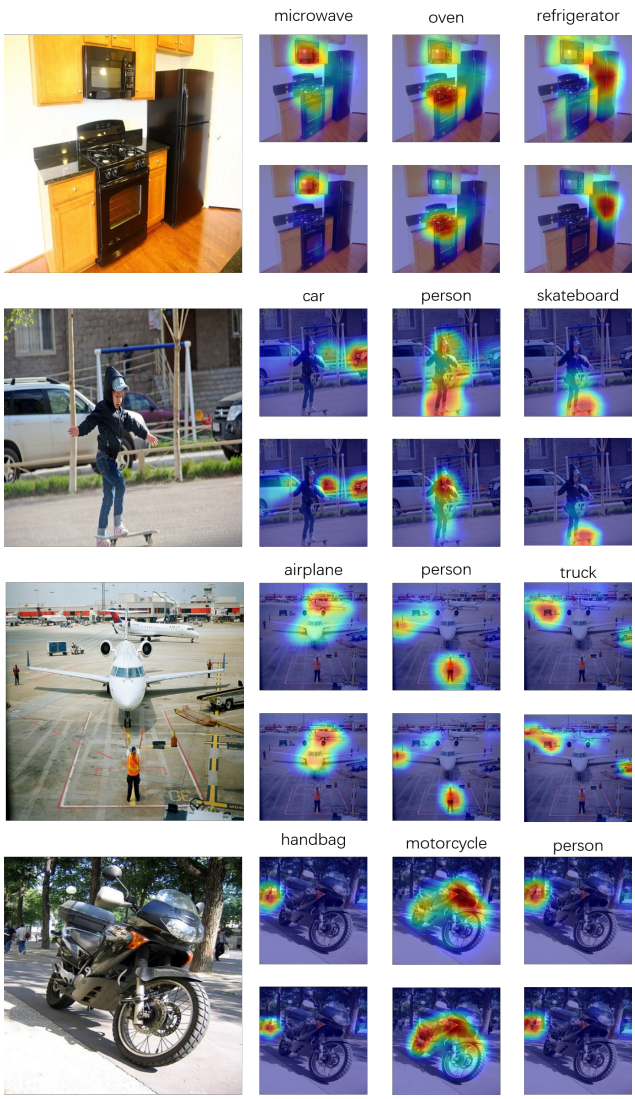


Figure 3. Heat maps on several images from the MS-COCO validation set. For each image, the top row is generated by the model trained by BN-BCE, and the bottom row is generated by the model trained by BN-BCE with SR. The object regions corresponding to present labels are highlighted in red.

involved in the training, while our method still complements many labels for them. We also find that the labels of some difficult objects (*e.g.*, handbag) are involved in the training, but the model still has difficulty identifying them.

Heat maps are presented on several images from the MS-COCO validation set in Fig. 3. The models trained by BN-BCE and BN-BCE with SR on the MS-COCO training set with 10% known labels generate these heat maps. As shown in Fig. 3, our SR focuses more on the object itself and less on the background region. The object regions corresponding to present labels are more salient after the addition of SR. SR enables the object regions to be focused more accu-

| Architectures | 10% | 30% | 50% | 70% | 90% |
|---|---|---|---|---|---|
| ResNet-101 | 77.2 | 80.3 | 81.8 | 82.1 | 82.7 |
| VIT-B16 | 81.2 | 83.9 | 84.8 | 85.2 | 85.4 |
| SWIN-B | 82.9 | 85.1 | 85.5 | 85.7 | 85.7 |
| EfficientNetV2-L | 83.1 | 86.3 | 87.3 | 87.8 | 87.9 |

Table 1. mAP of our method with different backbones on MS-COCO when different proportions of known labels.

| Methods | MS-COCO | VG-200 | VOC 2007 |
|---|---|---|---|
| S-BCE | 79.4 | 45.7 | 92.8 |
| SR | 82.5 | 50.2 | 94.3 |

Table 2. Comparison of mAP on different datasets in fully supervised mode.

rately, such as the microwave and oven of the first image. SR enables multiple object regions of a present label to be saliently focused, such as the car of the second image and the airplane of the third image. SR also enables different parts of the same object for a present label to be saliently focused, such as the motorcycle in the fourth image.

## F. Architecture Diversification

The current works on partial annotations are based on ResNet-101, so we use this backbone for a fair comparison in the main paper. To verify the effectiveness and generalizability of our method, we conduct experiments on some more sophisticated networks, including VIT-B16 [4], SWIN-B [7] and EfficientNetV2-L [9]. As shown in Tab. 1, compared with the performance on ResNet-101, Our method achieves better performance on these more sophisticated networks. Our method does not rely on a specific architecture and is a plug-and-play method.

## G. Wider Applications

We further explore the possibilities of our method for other tasks. The main paper discusses the multi-label classification task, so we expand the applications in its fully supervised mode with only saliency regularization (SR) and semi-supervised mode with self-training (ST).

**Fully supervised mode**. Our method works for the setting where the known labels are positive and negative. When the known labels are fully annotated, SR is still suitable for fully supervised learning. Since the unknown labels do not exist, consistency regularization (CR) loses its usefulness. In Tab. 2, the effectiveness of SR is verified on MS-COCO, VG-200 and VOC 2007. Compared with the performance of standard binary cross entropy (S-BCE), the performance of SR is significantly better.

**Semi-supervised mode**. In order to conduct semi-supervised experiments, we divide the dataset MS-COCO. According to different known proportions (*e.g.*, 10%, 30%, 50%, 70%, 90%), we randomly select a subset of images in the training set, and the images in this subset are fully annotated. None of the remaining images in the training set are annotated, and they participate in the model training together with the annotated images. Our method is suitable for semi-supervised learning, where SR is for supervised learning and CR is for unsupervised learning. When only supervised learning with known labels, the obtained performance is used as a baseline in Tab. 3. Compared with the baseline, the performance of our method is significantly better, which also verifies the effectiveness of our method in multi-label semi-supervised learning.

| Methods | 10% | 30% | 50% | 70% | 90% |
|---|---|---|---|---|---|
| Supervised | 66.7 | 74.5 | 76.9 | 77.9 | 78.1 |
| Ours | 72.6 | 77.9 | 79.4 | 80.1 | 80.5 |

Table 3. mAP comparison of our method and supervised-only learning in semi-supervised mode.

| Methods | G1 | G2 | G3 | G4 | G5 | All Gs |
|---|---|---|---|---|---|---|
| Latent Noise | 69.4 | 70.4 | 74.8 | 79.2 | 85.5 | 75.9 |
| CNN-RNN | 68.8 | 69.7 | 74.2 | 78.5 | 84.6 | 75.2 |
| Curriculum | 70.4 | 71.3 | 76.2 | 80.5 | 86.8 | 77.1 |
| IMCL | 71.0 | 72.6 | 77.6 | 81.8 | 87.3 | 78.1 |
| LL-Ct | 77.7 | 79.3 | 82.1 | 84.7 | 89.4 | 82.6 |
| CSL | 74.6 | 75.8 | 77.6 | 81.8 | 90.1 | 80.0 |
| CSL* | 73.2 | 78.6 | **85.1** | **87.7** | 90.6 | 83.0 |
| Ours | 76.0 | 77.7 | 79.5 | 83.1 | 91.2 | 81.5 |
| Ours* | **78.9** | **80.9** | 83.7 | 86.8 | **91.4** | **84.4** |

Table 4. Results on OpenImages V3. The 5000 categories are sorted in ascending order according to the number of available annotations for each category in the training set, and then divided into 5 groups on average, that is, each group (from G1 to G5) contains 1000 categories. All Gs contains all categories. The mAP score of each group is compared between our method and existing methods. The best results are marked in bold.

## H. Performance Details

The results in Tab. 4 are an extension of the results in Tab. 4 of the main paper. Due to the limited number of downloaded images in the OpenImages V3 training set, we only conduct experiments on a part of its training set (1.7M images) in the main paper. In Tab. 4, the reproduced CSL and Ours represent results of the main paper, and CSL* represents results of the original paper [1]. Meanwhile, we also introduce the state-of-the-art results of LL-Ct [6]. For a fair comparison, we also use the same training set (3.4M images) for experiments. Ours* shows the effectiveness of our method. It is observed that the performance gain from the extra 1.7M images is small.

In Fig. 3 of the main paper, for the performance comparison of our framework with other methods, we show the trend curves of mAP with various known proportions on the different datasets. To compare in more detail, we

present the specific mAP for different methods in the different known proportions on MS-COCO, VG-200, and VOC 2007 in Tab. 5. There are only the average mAP, OF1, and CF1 in Tab. 1 of the main paper. For a more comprehensive comparison, we introduce the average mAP, OP, OR, OF1, CP, CR, and CF1 in Tab. 6. We can find that the performance of our method continues to dominate in most metrics.

## References


[1] Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *CVPR*, pages 4764–4772, 2022. 3, 4

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 3

[3] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. Structured semantic transfer for multi-label recognition with partial labels. In *AAAI*, pages 339–346, 2022. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[5] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *CVPR*, pages 9423–9432, 2020. 3

[6] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *CVPR*, pages 14156–14165, 2022. 4

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 4

[8] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. In *AAAI*, pages 2091–2098, 2022. 2

[9] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106, 2021. 4

| Datasets | Methods | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | Ave. mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MS-COCO | SSGRL | 62.5 | 70.5 | 73.2 | 74.5 | 76.3 | 76.5 | 77.1 | 77.9 | 78.4 | 74.1 |
| | GCN-ML | 63.8 | 70.9 | 72.8 | 74.0 | 76.7 | 77.1 | 77.3 | 78.3 | 78.6 | 74.4 |
| | PLCL | 68.8 | 72.8 | 74.6 | 75.9 | 76.6 | 77.3 | 77.7 | 78.0 | 78.1 | 75.5 |
| | SPLC | 64.1 | 72.0 | 75.3 | 76.7 | 77.1 | 77.5 | 77.3 | 76.9 | 76.9 | 74.9 |
| | SST | 68.1 | 73.5 | 75.9 | 77.3 | 78.1 | 78.9 | 79.2 | 79.6 | 79.9 | 76.7 |
| | HST | 70.6 | 75.8 | 77.3 | 78.3 | 79.0 | 79.4 | 79.9 | 80.2 | 80.4 | 77.9 |
| | SARB | 71.2 | 75.0 | 77.1 | 78.3 | 78.9 | 79.6 | 79.8 | 80.5 | 80.5 | 77.9 |
| | SARB* | 72.5 | 76.0 | 77.6 | 78.7 | 79.6 | 79.8 | 80.0 | 80.5 | 80.8 | 78.4 |
| | CSL | 67.0 | 72.4 | 74.9 | 76.6 | 77.7 | 78.6 | 79.3 | 79.8 | 80.3 | 76.3 |
| | Ours | **77.2** | **79.2** | **80.3** | **80.9** | **81.8** | **82.1** | **82.1** | **82.6** | **82.7** | **81.0** |
| VG-200 | SSGRL | 34.6 | 37.3 | 39.2 | 40.1 | 40.4 | 41.0 | 41.3 | 41.6 | 42.1 | 39.7 |
| | GCN-ML | 32.0 | 37.8 | 38.8 | 39.1 | 39.6 | 40.0 | 41.9 | 42.3 | 42.5 | 39.3 |
| | PLCL | 40.5 | 42.9 | 43.9 | 44.5 | 44.9 | 45.0 | 45.1 | 45.2 | 45.3 | 44.1 |
| | SPLC | 39.5 | 42.8 | 45.0 | 45.9 | 46.4 | 46.6 | 46.6 | 46.5 | 46.3 | 45.1 |
| | SST | 38.8 | 39.4 | 41.1 | 41.8 | 42.7 | 42.9 | 43.0 | 43.2 | 43.5 | 41.8 |
| | HST | 40.6 | 41.6 | 43.3 | 44.6 | 45.2 | 45.8 | 46.8 | 47.2 | 47.8 | 44.8 |
| | SARB | 40.6 | 43.5 | 44.5 | 45.3 | 46.0 | 47.1 | 47.2 | 47.8 | 48.1 | 45.6 |
| | SARB* | 41.4 | 44.0 | 44.8 | 45.5 | 46.6 | 47.5 | 47.8 | 48.0 | 48.2 | 46.0 |
| | CSL | 40.7 | 43.7 | 45.2 | 46.2 | 46.8 | 47.3 | 47.8 | 48.2 | 48.5 | 46.0 |
| | Ours | **46.7** | **48.2** | **49.0** | **49.5** | **49.8** | **49.9** | **50.0** | **50.1** | **50.2** | **49.2** |
| VOC 2007 | SSGRL | 77.7 | 87.6 | 89.9 | 90.7 | 91.4 | 91.8 | 92.0 | 92.2 | 92.2 | 89.5 |
| | GCN-ML | 74.5 | 87.4 | 89.7 | 90.7 | 91.0 | 91.3 | 91.5 | 91.8 | 92.0 | 88.9 |
| | PLCL | 87.0 | 90.8 | 92.2 | 92.9 | 93.4 | 93.5 | 93.7 | 93.9 | 94.0 | 92.4 |
| | SPLC | 79.4 | 86.2 | 89.9 | 91.8 | 92.6 | 92.9 | 93.3 | 93.6 | 93.7 | 90.4 |
| | SST | 81.5 | 89.0 | 90.3 | 91.0 | 91.6 | 92.0 | 92.5 | 92.6 | 92.7 | 90.4 |
| | HST | 84.3 | 89.1 | 90.5 | 91.6 | 92.1 | 92.4 | 92.5 | 92.8 | 92.8 | 90.9 |
| | SARB | 83.5 | 88.6 | 90.7 | 91.4 | 91.9 | 92.2 | 92.6 | 92.8 | 92.9 | 90.7 |
| | SARB* | 85.7 | 89.8 | 91.8 | 92.0 | 92.3 | 92.7 | 92.9 | 93.1 | 93.2 | 91.5 |
| | CSL | 85.4 | 89.3 | 91.3 | 92.3 | 92.6 | 93.2 | 93.6 | 93.8 | 94.0 | 91.7 |
| | Ours | **91.3** | **93.1** | **93.6** | **94.0** | **94.1** | **94.3** | **94.3** | **94.3** | **94.3** | **93.7** |

Table 5. The detailed mAP of our ST framework and current SOTA methods for multi-label classification with partial labels at known labels of 10% to 90% on the MS-COCO, VG-200, and Pascal VOC 2007 datasets. The best results are marked in bold.

| Datasets | Methods | Avg. mAP | Avg. OP | Avg. OR | Avg. OF1 | Avg. CP | Avg. CR | Avg. CF1 |
|---|---|---|---|---|---|---|---|---|
| MS-COCO | SSGRL | 74.1 | 86.3 | 64.8 | 73.9 | 82.1 | 58.4 | 68.1 |
| | GCN-ML | 74.4 | 85.2 | 64.2 | 73.1 | 81.8 | 58.9 | 68.4 |
| | PLCL | 75.5 | 83.6 | 67.8 | 74.9 | 79.2 | 63.1 | 70.2 |
| | SPLC | 74.9 | 74.7 | 70.6 | 68.1 | 72.7 | 67.4 | 66.6 |
| | SST | 76.7 | 86.3 | 67.7 | 75.8 | 82.8 | 62.6 | 71.2 |
| | HST | 77.9 | - | - | 76.7 | - | - | 72.6 |
| | SARB | 77.9 | **86.6** | 68.6 | 76.5 | 82.9 | 64.1 | 72.2 |
| | SARB* | 78.4 | - | - | 76.8 | - | - | 72.7 |
| | CSL | 76.3 | 73.8 | **76.5** | 75.1 | 70.9 | **72.2** | 71.5 |
| | Ours | **81.0** | 85.3 | 73.5 | **79.0** | **83.3** | 69.4 | **75.7** |
| VG-200 | SSGRL | 39.7 | 69.9 | 25.9 | 37.8 | 45.3 | 18.3 | 26.1 |
| | GCN-ML | 39.3 | 64.1 | 28.2 | 38.7 | 44.6 | 18.2 | 25.6 |
| | PLCL | 44.1 | 65.8 | 35.1 | 45.8 | 55.3 | 30.5 | 39.3 |
| | SPLC | 45.1 | 57.0 | 48.3 | 43.9 | 49.6 | 43.0 | 41.1 |
| | SST | 41.8 | 69.9 | 27.9 | 39.9 | 49.8 | 22.3 | 30.8 |
| | HST | 44.8 | - | - | 46.3 | - | - | 37.9 |
| | SARB | 45.6 | **70.1** | 33.2 | 45.0 | 56.8 | 27.8 | 37.4 |
| | SARB* | 46.0 | - | - | 45.1 | - | - | 37.7 |
| | CSL | 46.0 | 54.8 | **53.2** | **54.0** | 48.9 | **47.1** | **48.0** |
| | Ours | **49.2** | 67.5 | 41.5 | 51.4 | **59.2** | 36.4 | 45.1 |
| VOC 2007 | SSGRL | 89.5 | 91.2 | 84.4 | 87.7 | 87.8 | 81.4 | 84.5 |
| | GCN-ML | 88.9 | 92.2 | 83.0 | 87.3 | 89.7 | 80.1 | 84.6 |
| | PLCL | 92.4 | 90.3 | 86.3 | 88.3 | 87.3 | 84.9 | 86.0 |
| | SPLC | 90.4 | 87.6 | 81.3 | 83.2 | 85.1 | 80.4 | 81.6 |
| | SST | 90.4 | 91.3 | 85.3 | 88.2 | 88.3 | 83.0 | 85.6 |
| | HST | 90.9 | - | - | 88.4 | - | - | 86.1 |
| | SARB | 90.7 | 93.0 | 83.6 | 88.4 | 90.4 | 81.1 | 85.9 |
| | SARB* | 91.5 | - | - | 88.3 | - | - | 86.0 |
| | CSL | 91.7 | 82.9 | **89.1** | 85.9 | 80.2 | **88.5** | 84.1 |
| | Ours | **93.7** | **93.2** | 85.1 | **88.9** | **91.5** | 81.6 | **86.2** |

Table 6. The average mAP, OP, OR, OF1, CP, CR, and CF1 of our ST framework and previous SOTA methods under the partial-label setting on the MS-COCO, VG-200, and Pascal VOC 2007 datasets. The best results are marked in bold. "-" denotes that the corresponding result is not provided.