Spatial Consistency Loss for Training
Multi-Label Classifiers from Single-Label Annotations

Supplementary material

A. Data-augmentation settings

We use the following data-augmentation pipeline during trainings:

MS-COCO 2014, Pascal VOC 2012, NUS-WIDE, CUB-Birds 200-2011

Train

- Resize to square image of resolution 672×672
- Random square crop with cropped area uniformly varying between 0.25 and 1 (torchvision [38] RandomResizedCrop implementation), resized to 448×448
- Random horizontal flip

Test

- Resize to square image of size 448×448

Imagenet-1k ILSVRC2012

Train

- Random square crop with cropped area uniformly varying between 0.08 and 1 and aspect ratio between 3/4 and 4/3 (torchvision [38] RandomResizedCrop implementation with default arguments, same as [19]), resized to size 224×224
- Random horizontal flip

Test

- Resize smallest image side to 256
- Center crop of 224×224 pixels

B. Comparison with Hill/SPLC

Zhang et al. [59] use different splits on MS-COCO [33] to evaluate training from a single positive label. In addition, they perform experiments on the partial label settings where 75% and 40% of the positive labels are annotated, and no annotated negatives. We evaluate our method on their dataset with our setup as described in section 4. Table B.1 shows that our results surpass those of [59] in all scenarios.

C. Ablation on the crop parameters

Figure B.1 shows the accuracies obtained with AN, and CL/SCL (with EN), when varying the random interval for the area of the crop data-augmentation. We see that CL and SCL are able to benefit more from the crop data-augmentation, compared to AN. This is consistent with our intuition that the crop data-augmentation can lead to incorrect supervision due to the single annotated objects being possibly partially or entirely cropped out. Moreover, SCL’s improvements over CL are consistent over the different data-augmentation parameters.

D. Analysis over object sizes

We check the impact of object size by splitting the positive annotations of the COCO val split into equally-sized bins, grouped by relative area of the ground truth bounding box. Then, for each bin we compute the mAP using the positive labels within that bin, and negatives over the whole val split since negatives have no object size. Figure D.1 shows that the usage of consistency loss (CL) and spatial consistency loss (SCL) both improve mAP for all object sizes, compared to the AN baseline. Interestingly, SCL
Table B.1. Comparison with Hill/SPLC \cite{59} with ResNet-50 \cite{19} on MS-COCO \cite{33}. Results with † are reported by \cite{59}.

| Baselines | 75% labels | 40% labels | 1 label |
|-----------|------------|------------|---------|
| BCE (fully annotated) | 80.32 | 80.32 | 80.32 |
| AN † | 76.81 | 70.49 | 68.57 |
| WAN † | 77.25 | 72.05 | 70.17 |
| BCE-LS † | 78.27 | 73.13 | 70.53 |
| Focal \cite{32} † | 76.95 | 71.66 | 70.19 |
| ASL \cite{42} † | 77.97 | 72.70 | 71.67 |
| Hill \cite{59} † | 78.84 | 75.15 | 73.17 |
| BCE + pseudo label † | 77.05 | 71.46 | 69.77 |
| ROLE \cite{9} † | 78.43 | 73.67 | 70.90 |
| Focal margin + SPLC \cite{59} † | 78.44 | 75.69 | 73.18 |
| Ours | BCE (fully annotated) | 80.2 | 80.2 | 80.2 |
| AN | 76.8 | 71.7 | 69.8 |
| EN + CL | 77.6 | 75.8 | 74.1 |
| EN + SCL | 79.3 | 75.9 | 74.7 |

![Figure E.1](image.png)

Figure E.1. Analysis of distance functions with $\ell_1$ norm, $\ell_2$ norm, Jensen–Shannon divergence (JSD) over weights $\gamma$.

yields higher mAP gains for smaller object sizes. Our hypothesis is that smaller objects are more likely to be cropped out, which is handled by the SCL. In addition, the crop augmentation zooms in on small objects, and those soft labels are recorded in the heatmaps as supervision.

E. Ablation of distance functions and weights

Figure E.1 compares different distance functions to measure the difference between exponential moving averages and predictions for (spatial) consistency losses.

F. Score distributions

Figure F.1 shows the distributions of the top-4 scores over all validation images. In contrast to the fully annotated baseline, the single-positive dataset in combination with AN loss leads to low-scoring predictions. SCL with EN loss (eq. (8)) reduces the amount of false negative labels and leads to a distribution more akin to the fully annotated case.

G. Details on heatmaps computation

We store heatmaps on 2 times the resolution of the feature maps (e.g. input resolution of $448 \times 448$ results in feature maps of $14 \times 14$ is stored in heatmaps of $28 \times 28$ pixels.). Heatmaps are stored in 8-bit unsigned integer format.

For ImageNet-1K \cite{10} (section 4.3), we reduce the memory load by only keeping heatmaps for the top-$k$ classes. The selection is based on the per-class EMA scores $s_{ni}$ computed as described in eq. (4), after the 5 epochs of pretraining the linear layer. In our experiments, we select the 10 highest-scoring classes per image based on $s_{ni}$. Heatmaps of other classes are assumed to be uniformly 0 in the SCL. Given 1.3 million training images, heatmaps of $14 \times 14$ and 1000 classes stored in uint8, this optimization reduces the required memory from approximately 250 GB to 2.5 GB.

H. Impact of SCL on heatmaps

A comparison of the heatmaps generated with and without SCL is given in fig. H.1 as an extra example in addition to fig. 4.

I. Uncurated heatmap examples

Figures K.1 and K.2 show the heatmaps corresponding to the samples with lowest COCO image id having suitable licenses for reproduction in the paper. In agreement with the observations in section 4.2 we see that the SCL tends to improve the object localization in the heatmaps, especially.
Figure F.1. Score distribution over all MS-COCO validation images, for 1st, 2nd, 3rd and 4th highest predicted scores per image. The BCE method is a fully annotated baseline. Training with AN and a single-positive label leads to a bias towards single positive predictions. With EN and SCL, the network more confidently predicts multiple positives.

Figure H.1. Comparison of heatmaps generated in the final training epoch with and without spatial consistency loss (second example).

when looking at the negative classes which tend to be more present when using the EN alone.

**J. Distribute property of final pooling and linear layer**

To obtain predictions for each spatial position, we flip the order of the average pooling layer and the final linear classification layer. The linear layer can be executed as a $1 \times 1$ convolution over the feature map, resulting in class-wise predictions per spatial position. While this introduces extra computations at training time, the inference time is not impacted. Due to the distributive property, the order of the average pooling and $1 \times 1$ convolutions can be reversed at inference time without affecting the network outputs. Indeed, denoting by $\phi$ the $G \times G \times M$ network output before average pooling and $1 \times 1$ convolution, and by $A$ the $M \times L$ matrix representing the $1 \times 1$ convolution, it can be seen that

$$\frac{1}{G^2} \sum_{g,g'=1}^{G} \sum_{m=1}^{M} A_{ml} \phi_{g'g} m = \sum_{m=1}^{M} A_{ml} \frac{1}{G^2} \sum_{g,g'=1}^{G} \phi_{gg'} m$$

for all $l$. That is, convolving and then average pooling is equal to average pooling and then convolving.

**K. Dataset statistics**

Table K.1 lists some statistics on the datasets used in the paper, as well as the value of the hyperparameter $K$ computed on the validation set based on these statistics. Tables K.2 and K.3 show detailed breakdown of positive annotations per class in the MS-COCO and Pascal datasets using the splits of [?].
Figure K.1. Heatmaps and scores of the top-5 scoring classes in the last epoch training with EN+SCL, along with the corresponding heatmaps for EN alone.
Figure K.2. Heatmaps and scores of the top-5 scoring classes in the last epoch training with EN+SCL, along with the corresponding heatmaps for EN alone.
Table K.1. Dataset statistics. For COCO, VOC, NUS and CUB we use the train/val/test splits from \cite{lin2014microsoft}. For ImageNet-1K we report both the original \cite{deng2009imagenet} and multi-label ReaL \cite{goyal2017una} validation sets. $K$ is the average number of positives per image on the validation set.

| Dataset           | Num. classes | Number of images | Number of annotations | $K$  |
|-------------------|--------------|------------------|----------------------|------|
|                   |              | train | val | test | train | val | test |      |
| MS-COCO 2014 \cite{lin2014microsoft} | 80       | 65,665| 16,416| 40,137 | 193078| 47957| 116592| 2.9  |
| Pascal VOC 2012 \cite{everingham2010pascal} | 20       | 4574 | 1143 | 5823 | 6665 | 1143 | 5823 | 1.5  |
| NUS-WIDE \cite{chen2015nus} | 81       | 120000| 30000| 60260 | 226833| 57778| 113418| 1.9  |
| CUB-200-2011 \cite{wah2011caltech} | 312      | 4795 | 1199 | 5794 | 150551| 37792| 182704| 31.5 |
| ImageNet-1K \cite{deng2009imagenet} | 1000     | 1,281,167| 50,000/46,837 | - | 1,281,167| 50,000/46,837 | - | 1/1.2 |
Table K.2. Annotation statistics on MS-COCO [33]. For each class, we show the total amount of annotations in the original MS-COCO annotations (total), as well as the percentage of single-positive annotations selected for that class in the splits of [9].

| Class     | # train | # val | # test |
|-----------|---------|-------|--------|
| all classes | 193078 | 47957 | 116592 |
| person    | 36192  | 34%   | 8982   |
| chair      | 7138   | 22%   | 1812   |
| car        | 6895   | 30%   | 1711   |
| dining table | 6701 | 21%   | 1677   |
| cup        | 5219   | 20%   | 1299   |
| bottle     | 4790   | 20%   | 1178   |
| bowl       | 4042   | 21%   | 986    |
| handbag    | 3927   | 23%   | 934    |
| truck      | 3447   | 33%   | 874    |
| backpack  | 3109   | 25%   | 815    |
| bench      | 3078   | 34%   | 766    |
| book       | 2994   | 22%   | 746    |
| cell phone | 2644   | 29%   | 678    |
| sink       | 2640   | 33%   | 651    |
| tv         | 2525   | 23%   | 666    |
| couch      | 2515   | 22%   | 655    |
| clock      | 2506   | 50%   | 655    |
| potted plant | 2497 | 24%   | 587    |
| knife      | 2491   | 20%   | 609    |
| dog        | 2428   | 39%   | 613    |
| sports ball | 2401 | 30%   | 585    |
| traffic light | 2292 | 37%   | 601    |
| cat        | 2267   | 43%   | 551    |
| bus        | 2240   | 33%   | 551    |
| umbrella   | 2183   | 30%   | 566    |
| tie        | 2132   | 34%   | 535    |
| fork       | 2058   | 18%   | 479    |
| bed        | 2054   | 48%   | 385    |
| vase       | 2025   | 35%   | 505    |
| skateboard | 2021   | 40%   | 490    |
| spoon      | 2005   | 18%   | 488    |
| motorcycle | 1961   | 37%   | 481    |
| train      | 1958   | 58%   | 506    |
| laptop     | 1943   | 24%   | 532    |
| tennis racket | 1903 | 35%   | 465    |
| surfboard  | 1876   | 44%   | 467    |
| toilet     | 1842   | 58%   | 475    |
| airplane   | 1797   | 68%   | 446    |
| bird       | 1784   | 64%   | 457    |

Table K.3. Annotation statistics on Pascal VOC 2012 [14]. For each class, we show the total amount of annotations in the original MS-COCO annotations (total), as well as the percentage of single-positive annotations selected for that class in the splits of [9].

| Class     | # train | # val | # test |
|-----------|---------|-------|--------|
| all classes | 6665   | 68%   | 1666   |
| person    | 1584   | 59%   | 410    |
| dog       | 504    | 83%   | 128    |
| car       | 474    | 68%   | 116    |
| chair     | 457    | 49%   | 107    |
| cat       | 436    | 90%   | 103    |
| bird      | 310    | 93%   | 85     |
| bottle    | 294    | 51%   | 71     |
| aeroplane | 264    | 95%   | 63     |
| ryder     | 233    | 61%   | 57     |

| Class     | # train | # val | # test |
|-----------|---------|-------|--------|
| train     | 220    | 85%   | 53     |
| pottedplant | 214 | 47%   | 35     |
| motorbike | 206    | 65%   | 59     |
| sofa      | 201    | 53%   | 56     |
| bicycle   | 200    | 64%   | 68     |
| horse     | 195    | 69%   | 42     |
| bus       | 176    | 67%   | 37     |
| sheep     | 135    | 90%   | 36     |
| cow       | 129    | 86%   | 22     |