南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

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

WACV 2023

Problem to be solved : single positive label

	person	dog	bus	bicycle	apple	boat	laptop	couch
(a)	✓	×	✓	✓	×	×	×	×
(b)	✓	×	?	✓	?	×	?	?
(c)	✓	?	?	?	?	?	?	?

(a) full annotations

(b) partial annotations

(c) single positive label

Method

Multi-label classification with Partially annotated labels

dataset $(\mathbf{x}_n, \mathbf{z}_n)_{n=1}^{N}$

$\mathbf{z}_n \in \{0, 1, \emptyset\}^L$ negative, positive, unknown

Method pipeline

The spatial consistency between network's out and EMAs out

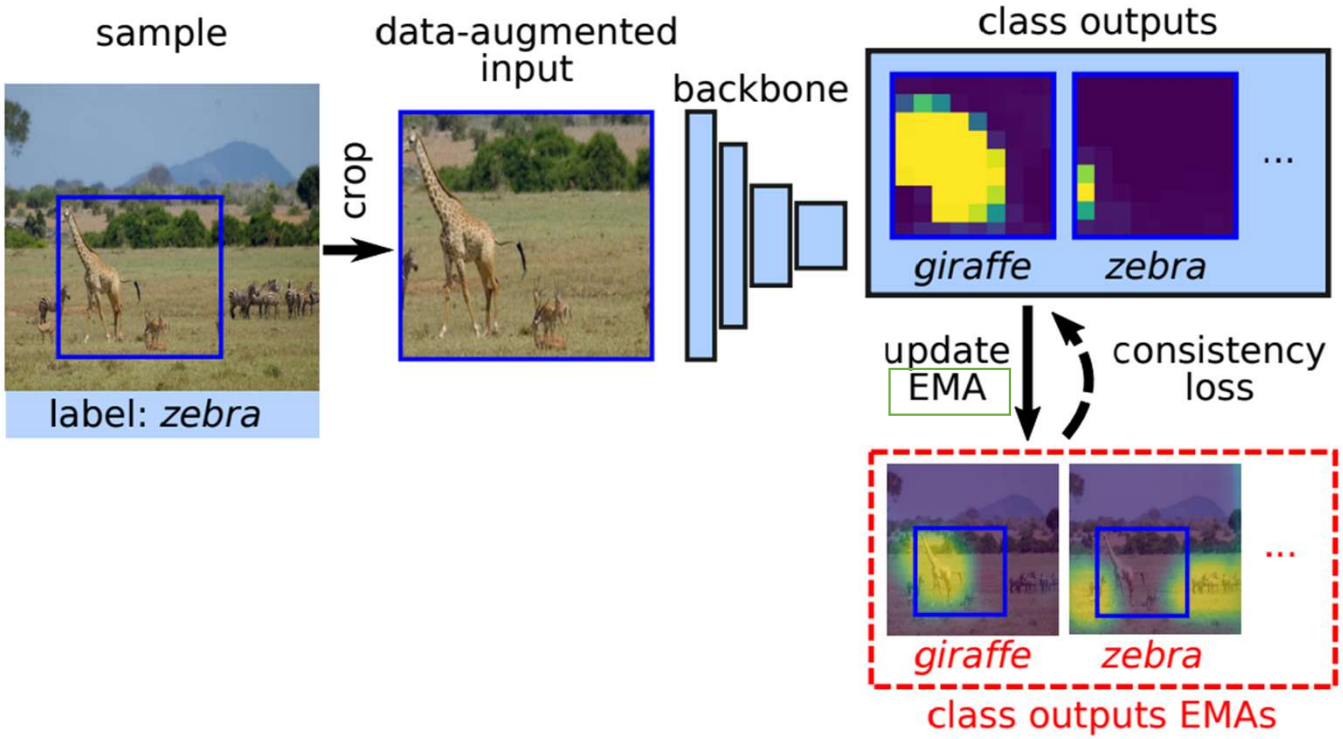


Photo by S. Cozens

Method : Assume negative

BCE

$$\mathcal{L}_{\mathrm{BCE}}(\mathbf{f}_n) = -\frac{1}{L}\sum_{i=1}^{L}[z_{ni}=1]\log(f_{ni}) + [z_{ni}=0]\log(1-f_{ni}) \quad (1)$$

AN : assume that all unknown labels are negatives

$$\mathcal{L}_{\mathrm{AN}}(\mathbf{f}_n) = -\frac{1}{L}\sum_{i=1}^{L}[z_{ni}=1]\log(f_{ni}) + [z_{ni}\in\{0,\emptyset\}]\log(1-f_{ni})\,. \quad (2)$$

Method : Expected negative

ignore the large incorrect contributions of noisy labels,
Top-p_i,

$$p_i = KN \cdot \frac{\sum_{n=1}^{N} [z_{ni} = 1]}{N} = K \sum_{n=1}^{N} [z_{ni} = 1], \quad (3)$$

$$\sum_{n=1}^{N} [z_{ni} = 1]/N \longrightarrow \sum_{n=1}^{N} y_{ni}/N. \qquad \hat{z}_{ni}^{t} \in \{0, 1\}$$

Annotated label class distribution

true label class distribution

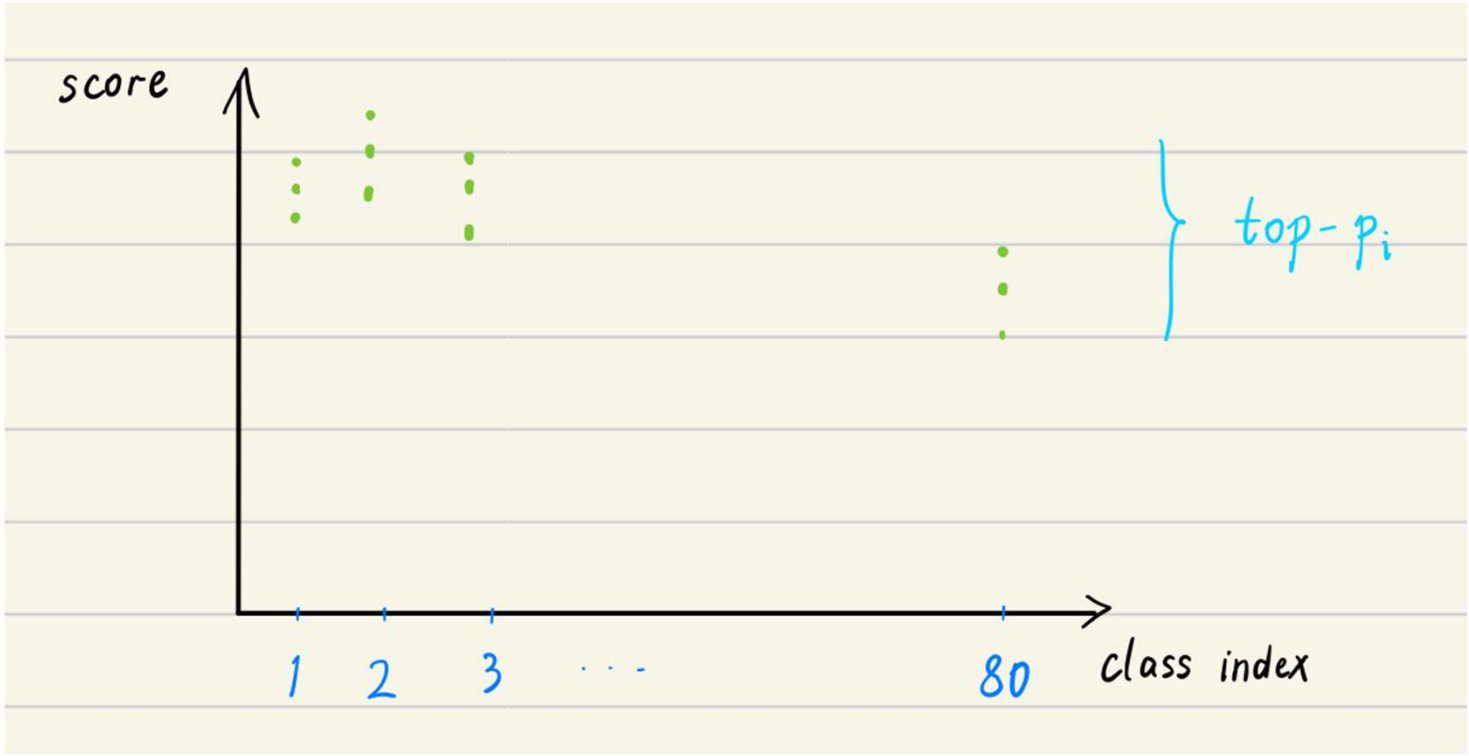
estimated score **s** are updated with network output **f**

$$\mathbf{s}_n^t = \mu\, \mathbf{s}_n^{t-1} + (1 - \mu)\mathbf{f}_n^t \quad (4)$$

Method : Expected negative

Top p_i instances per image for a class

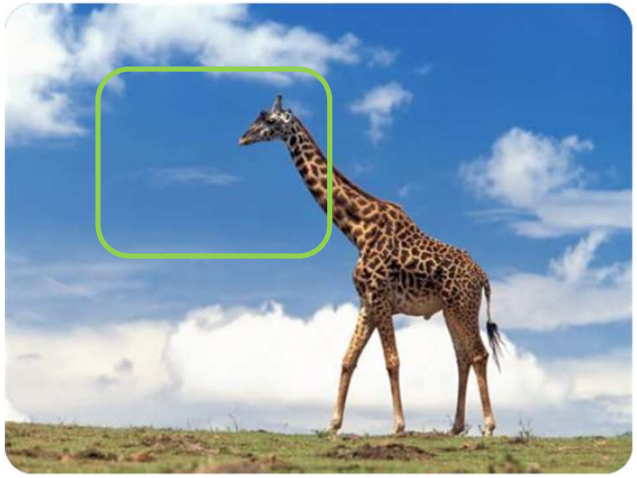
$$\hat{z}_{ni}^{t} \in \{0, 1\}$$



Method : Expected negative

$$p_i = KN \cdot \frac{\sum_{n=1}^{N} [z_{ni} = 1]}{N} = K \sum_{n=1}^{N} [z_{ni} = 1], \quad (3)$$

$$\mathbf{s}_n^t = \mu\, \mathbf{s}_n^{t-1} + (1-\mu)\mathbf{f}_n^t \quad (4)$$



But label drift, … …

Only annotated positives, expected negative, ignore expected positive

$$\mathcal{L}_{\mathrm{EN}}(\mathbf{f}_n) = -\frac{1}{L} \sum_{i=1}^{L} [z_{ni}{=}1] \log(f_{ni}) + [\hat{z}_{ni}^t{=}0] \log(1{-}f_{ni}). \quad (5)$$

Method : Consistency loss

Use robust targets from AN as supervisions

$$\mathcal{L}_{\mathrm{CL}}(\mathbf{f}_n^t) = \|\mathbf{f}_n^t - \mathbf{s}_n^{t-1}\|_1 . \qquad (6)$$

Method : Spatial Consistency loss

New noisy label, as object can be crop out,
So use the running average in spatial dimension score heatmap.

L1-distance (score heatmap, network output)

$$\mathcal{L}_{\text{SCL}}(\mathbf{F}_n^t) = \|\mathbf{F}_n^t - \text{resize}(T_n^t(\mathbf{H}_n^{t-1}))\|_1 \,. \qquad (7)$$

$$\mathcal{L} = \mathcal{L}_{\text{EN}} + \gamma\mathcal{L}_{\text{(S)CL}} \quad . \qquad (8)$$

Experiments

	Method	Supervision	No pretraining		IN1K pretraining			
			VOC12	MS-COCO	VOC12	MS-COCO	NUS	CUB
	fully-annotated oracle (BCE)	all pos + all neg	53.1	66.1	90.0	79.4	53.7	33.2
Related work	AN + label smoothing [9]†	1 pos / img	-	-	86.5	69.2	44.9	17.9
	ROLE (reported in [9])†	1 pos / img	-	-	88.2	69.0	**51.0**	16.8
	LL-R (reported in [25])†	1 pos / img	-	-	**89.4**	71.9	49.1	21.5
	LL-Ct (reported in [25])†	1 pos / img	-	-	89.3	71.6	49.6	21.8
	LL-Cp (reported in [25])†	1 pos / img	-	-	89.3	71.0	49.4	21.4
Baselines	Assume negative (AN)	1 pos / img	46.5	49.1	86.0	69.0	45.5	21.1
	AN + label smoothing	1 pos / img	46.0	46.1	87.6	70.3	46.7	16.0
	WAN [9] (our training schedule)	1 pos / img	44.4	45.1	86.4	69.3	45.6	21.3
	ROLE [9] (our training schedule)	1 pos / img	45.0	51.9	87.8	69.9	47.8	20.3
Ours	Expected Negative (EN)	1 pos / img	47.5	53.4	88.1	71.8	49.1	22.3
	EN + consistency loss (CL)	1 pos / img	49.1	**55.0**	88.3	71.9	49.0	22.1
	EN + spatial consistency (SCL)	1 pos / img	**51.4**	54.0	88.8	**73.2**	50.3	**22.5**

Table 1. Mean average precision (mAP) obtained on the test set of Pascal VOC 2012 [14] and MS-COCO 2014 [33], NUS-WIDE [8] and CUB [49]. ImageNet-1K [43] pretraining warms up the linear layer for 5 epochs. Results indicated with † are reported by related work.

Revisit to the structure of CNN

$$\mathbf{F} = [F_1, F_2, ..., F_C] \in \mathbb{R}^{C \times W \times H}$$

channel dimension

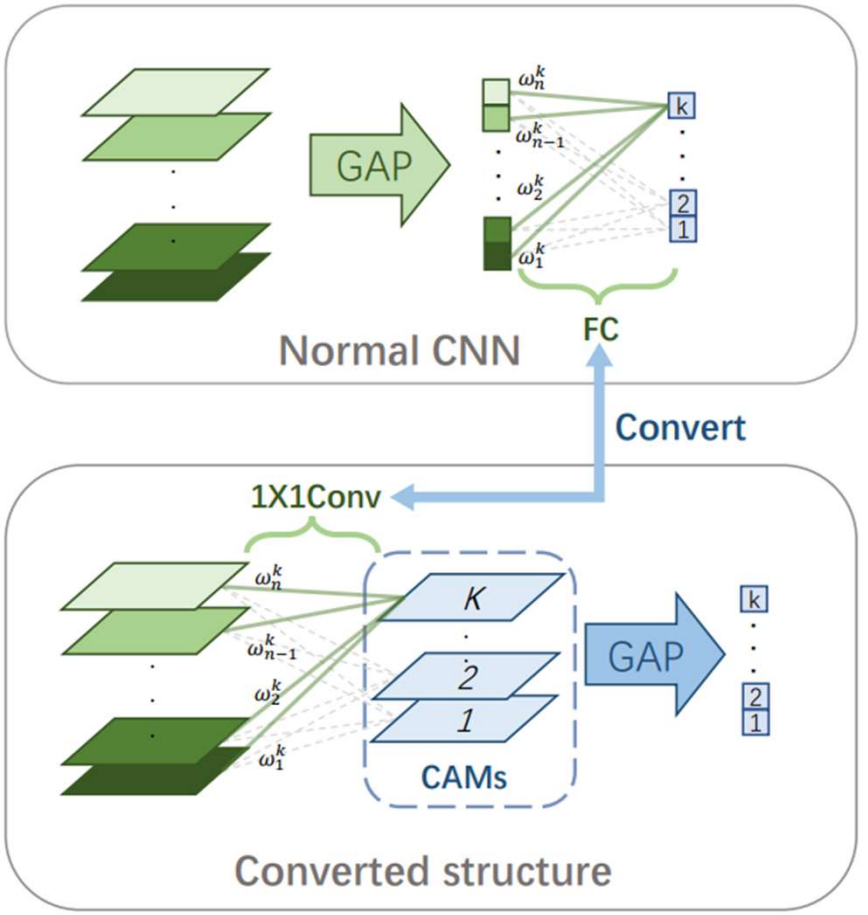


Figure 1. Illustration of the converted structure. After converting the FC layer into a convolutional layer with 1×1 kernel and moving the position of the global average pooling layer, CAMs can be obtained during the forward propagation.

Revisit to the structure of CNN

$$\mathbf{F} = [F_1, F_2, ..., F_C] \in \mathbb{R}^{C \times W \times H}$$

Classification logits from CNN :

$$L_i = \sum_{1 \le j \le C} \omega_j^i \times GAP(F_j)$$

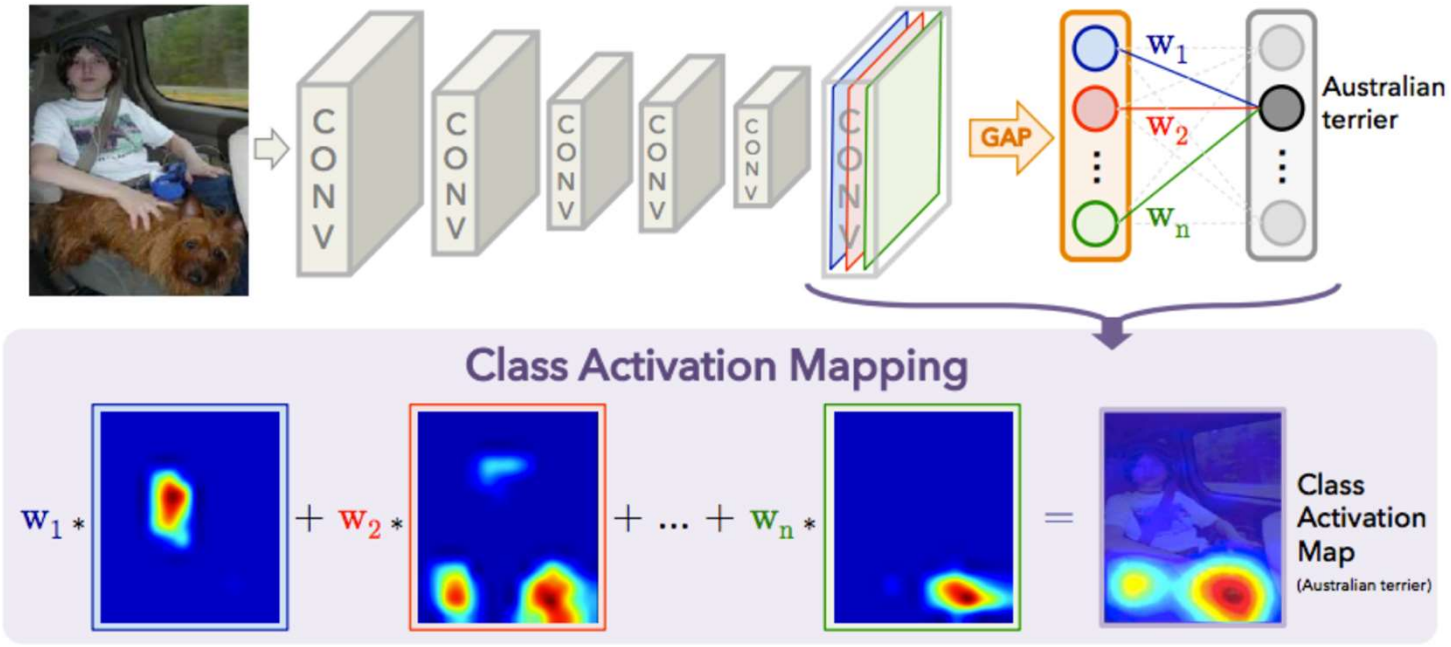
class i　　channel j

$$= \frac{1}{W \times H} \sum_{x,y} \sum_{1 \le j \le C} \omega_j^i \times f_j(x, y), \qquad (12.1)$$

activation of F

Revisit to the structure of CNN

From ‘Learning Deep Features for Discriminative Localization’



$$CAM_i(x,y) = \sum_{1 \le j \le C} \omega_j^i \times f_j(x,y). \quad (12.2)$$

Revisit to the structure of CNN

From ‘Learning Deep Features for Discriminative Localization’

$$L_i = \sum_{1 \le j \le C} \omega_j^i \times GAP(F_j)$$

$$= \frac{1}{W \times H} \sum_{x,y} \sum_{1 \le j \le C} \omega_j^i \times f_j(x,y), \qquad (12.1)$$

activation of F

$$CAM_i(x,y) = \sum_{1 \le j \le C} \omega_j^i \times f_j(x,y). \qquad (12.2)$$

$$L_i = \frac{1}{W \times H} \sum_{x,y} CAM_i(x,y) \qquad (12.3)$$

average activation

$$= GAP(CAM_i).$$

Revisit to the structure of CNN

From ‘Learning Deep Features for Discriminative

$$L_i = \frac{1}{W \times H} \sum_{x,y} CAM_i(x, y)$$
$$= GAP(CAM_i).$$

The same

(12.3)

$$\bar{L}_i = GAP(Conv_i(\mathbf{F}))$$
$$= \frac{1}{W \times H} \sum_{x,y} \Big(\sum_{1 \le j \le C} \omega_j^i \times f_j(x, y) \Big)$$
$$= GAP(CAM_i),$$

(12.4)

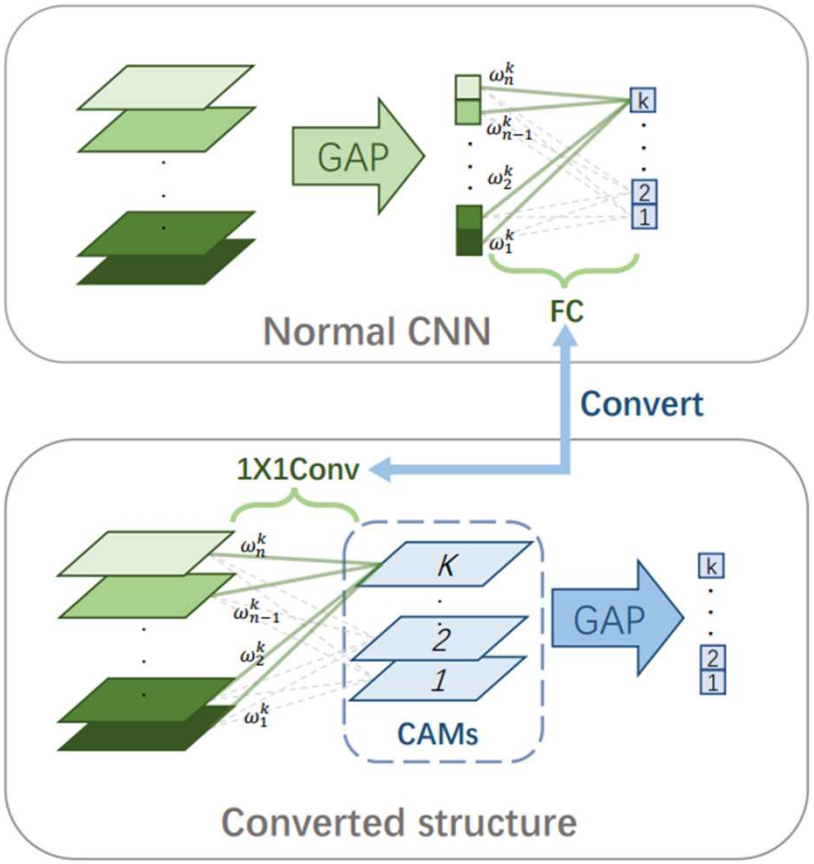


Figure 1. Illustration of the converted structure. After converting the FC layer into a convolutional layer with 1×1 kernel and moving the position of the global average pooling layer, CAMs can be obtained during the forward propagation.

Analysis and ablation : Spatial heatmaps

The top-k class

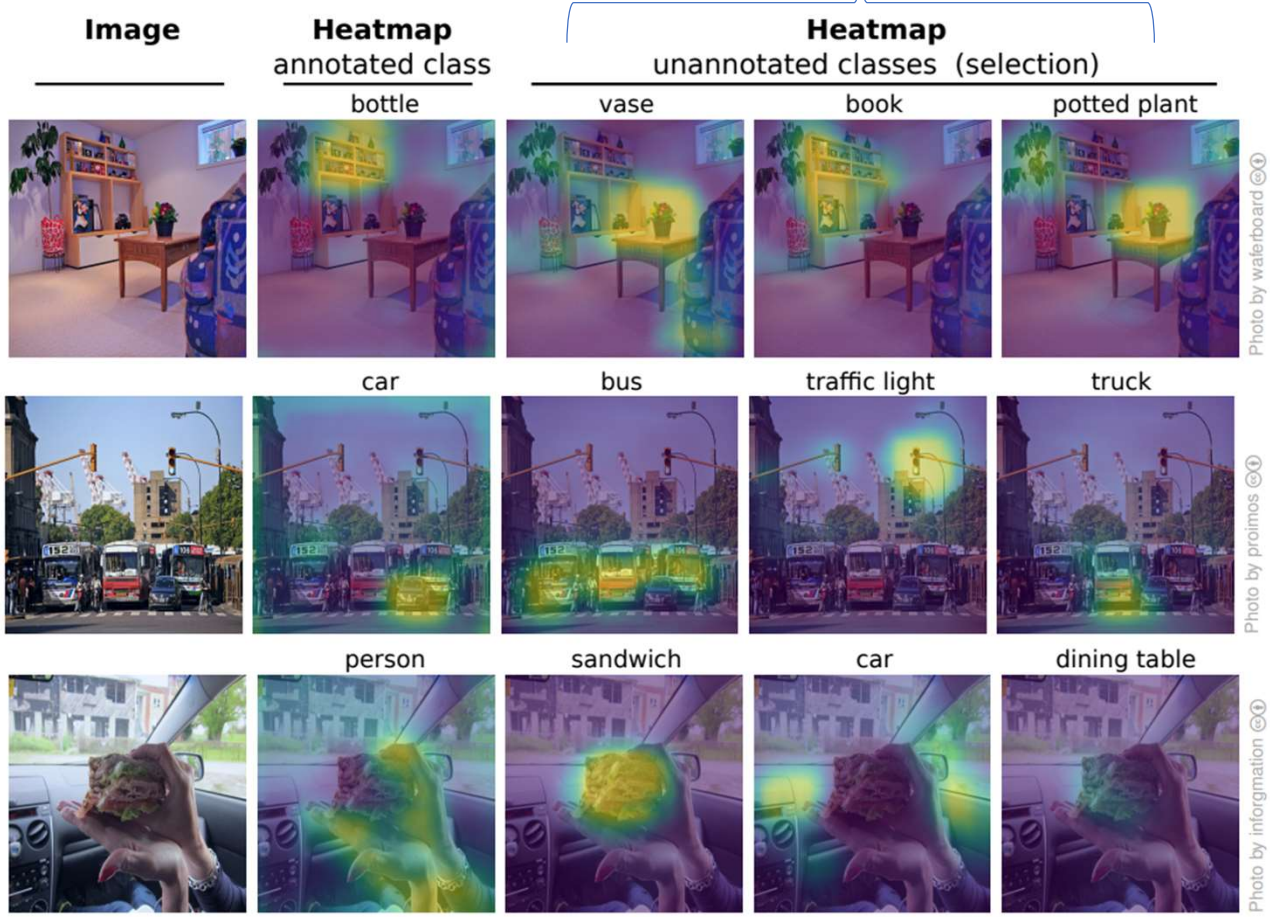


Figure 2. Heatmaps produced by ResNet-50 on MS-COCO in the last training epoch, with ImageNet pretraining (best viewed in color).

Localization of many objects in the image absent from the single-label ground truth.

Analysis and ablation : Spatial heatmaps

Process from training

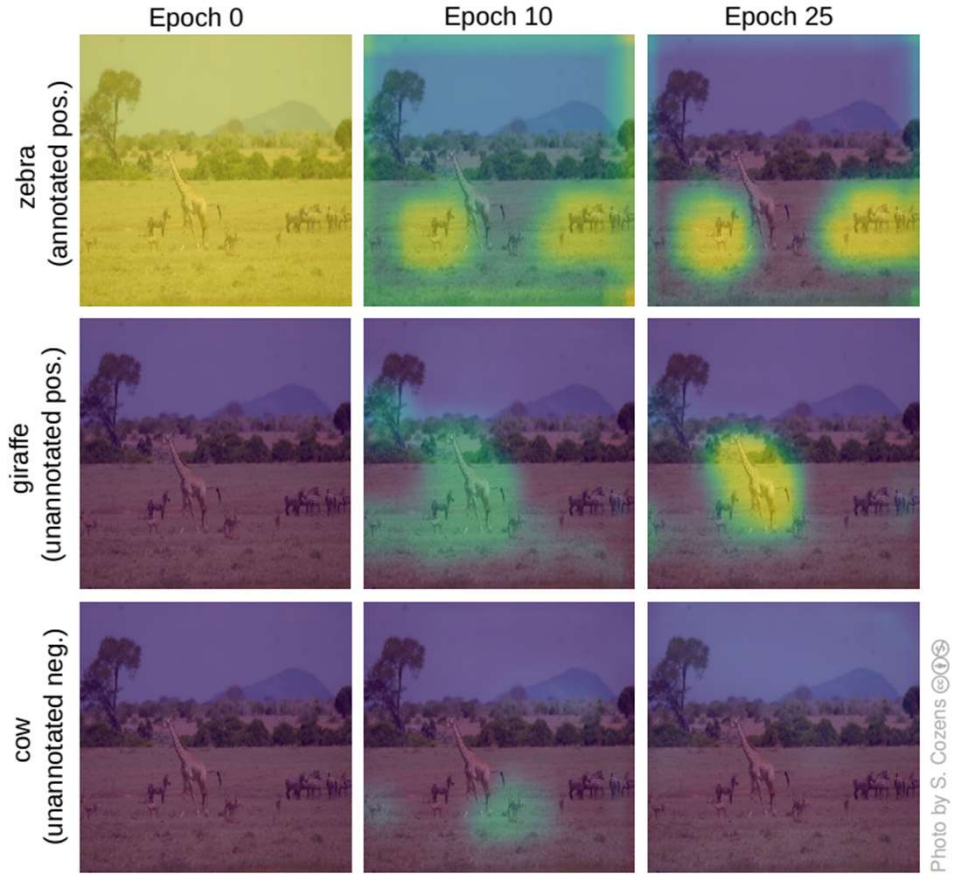


Figure 3. Progress of running-average heatmaps during training for an annotated positive class, unannotated positive class and negative class (best viewed in color).

Analysis and ablation : Spatial heatmaps

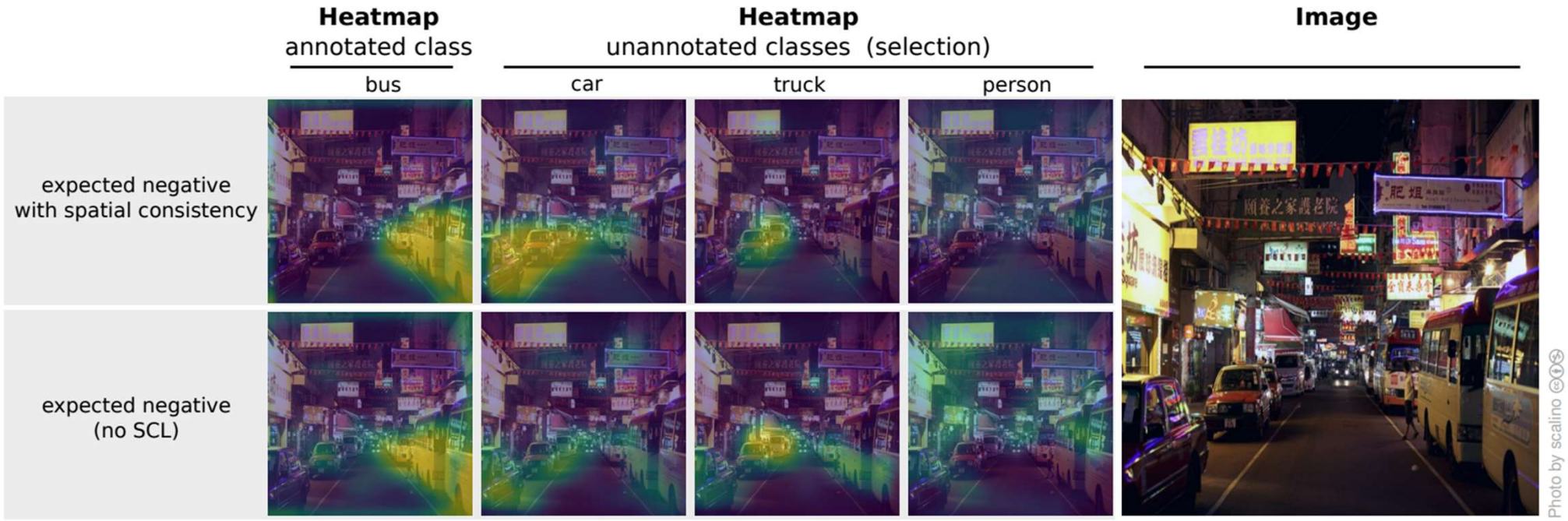
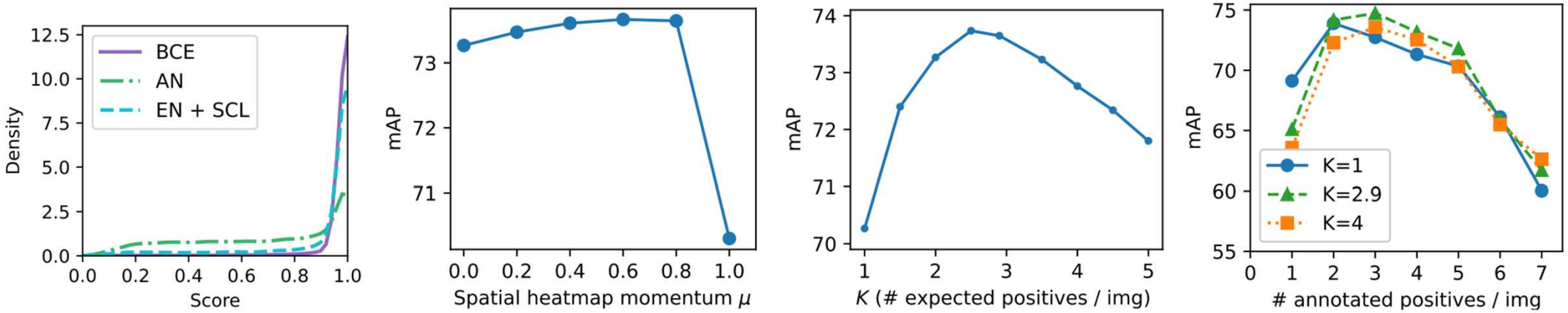


Figure 4. Comparison of heatmaps generated in the final training epoch with and without spatial consistency loss.

SCL localizes objects more precisely, avoiding false predictions for negative classes.

Analysis and ablation : Bias towards single-positive predictions



(a) Top-1 score distribution (b) momentum μ (c) hyperparameter K (d) # annot. positives score bins

Figure 5. Ablations on MS-COCO validation set with ImageNet-pretrained ResNet-50.

Fig 5(a). Distribution of the top-1 scores, per method, over all validation images.

Analysis and ablation : Bias towards single-positive predictions

Method	Loss	mAP
assume negative (AN)	$\mathcal{L}_{\text{AN}}$	69.4
expected negative (EN)	$\mathcal{L}_{\text{EN}}$	72.3
assume negative + CL	$\mathcal{L}_{\text{AN}} + \mathcal{L}_{\text{CL}}$	70.1
expected negatives + CL	$\mathcal{L}_{\text{EN}} + \mathcal{L}_{\text{CL}}$	**72.4**
expected positives and neg. + CL	$\mathcal{L}_{\text{EP}} + \mathcal{L}_{\text{CL}}$	65.8
expected positive regression [9] + CL	$\mathcal{L}_{\text{EPR}}$ [9] + $\mathcal{L}_{CL}$	71.7
assume negative + SCL	$\mathcal{L}_{\text{AN}} + \mathcal{L}_{\text{SCL}}$	70.2
expected negatives + SCL	$\mathcal{L}_{\text{EN}} + \mathcal{L}_{\text{SCL}}$	**73.7**
expected positives and neg. + SCL	$\mathcal{L}_{\text{EP}} + \mathcal{L}_{\text{SCL}}$	64.6
expected positive regression [9] + SCL	$\mathcal{L}_{\text{EPR}}$ [9] + $\mathcal{L}_{SCL}$	72.3

Table 2. Methods to avoid single-pos. bias (MS-COCO *val* split).

EN loss ignores expected positive samples.

EP (expected positive) use expected positive as additional positives in the supervision; (incorrect expected-positives)

EPR (regression) regresses the sum of the predicted probabilities towards the estimated number of positive K.

Analysis and ablation : Bias towards single-positive predictions

Method	Loss	mAP
assume negative (AN)	$\mathcal{L}_{\text{AN}}$	69.4
expected negative (EN)	$\mathcal{L}_{\text{EN}}$	72.3
assume negative + CL	$\mathcal{L}_{\text{AN}} + \mathcal{L}_{\text{CL}}$	70.1
expected negatives + CL	$\mathcal{L}_{\text{EN}} + \mathcal{L}_{\text{CL}}$	**72.4**
expected positives and neg. + CL	$\mathcal{L}_{\text{EP}} + \mathcal{L}_{\text{CL}}$	65.8
expected positive regression [9] + CL	$\mathcal{L}_{\text{EPR}}$ [9] + $\mathcal{L}_{CL}$	71.7
assume negative + SCL	$\mathcal{L}_{\text{AN}} + \mathcal{L}_{\text{SCL}}$	70.2
expected negatives + SCL	$\mathcal{L}_{\text{EN}} + \mathcal{L}_{\text{SCL}}$	**73.7**
expected positives and neg. + SCL	$\mathcal{L}_{\text{EP}} + \mathcal{L}_{\text{SCL}}$	64.6
expected positive regression [9] + SCL	$\mathcal{L}_{\text{EPR}}$ [9] + $\mathcal{L}_{SCL}$	72.3

Table 2. Methods to avoid single-pos. bias (MS-COCO *val* split).

EN loss ignores expected positive samples.

EP (expected positive) use expected positive as additional positives in the supervision; (incorrect expected-positives)

EPR (regression) regresses the sum of the predicted probabilities towards the estimated number of positive K.

Ablation on the crop parameter

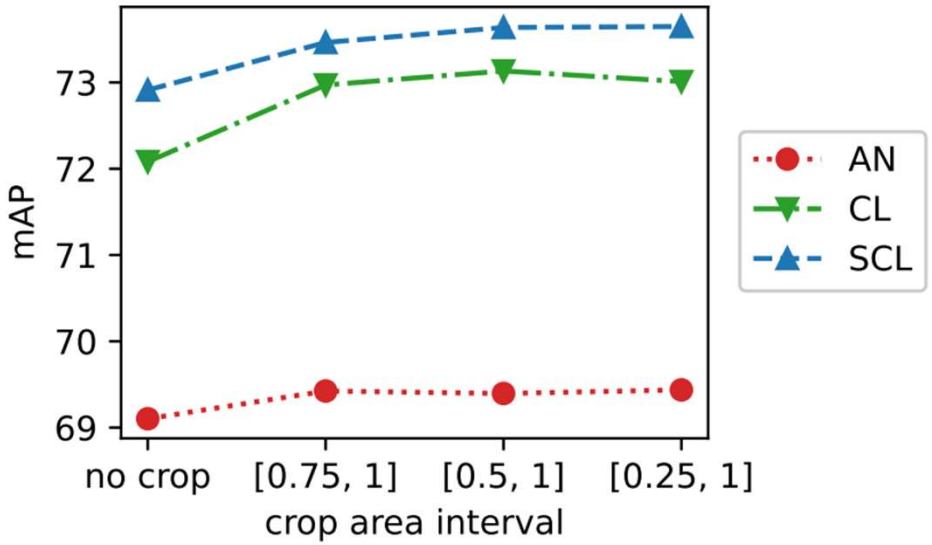


Figure B.1. Best MS-COCO validation mAP obtained when training with different data-augmentation crop area. The cropped area size, compared to the full image area, is randomly and uniformly sampled from the interval.

varying the random interval for the area of the crop data-augmentation.

Analysis over object sizes

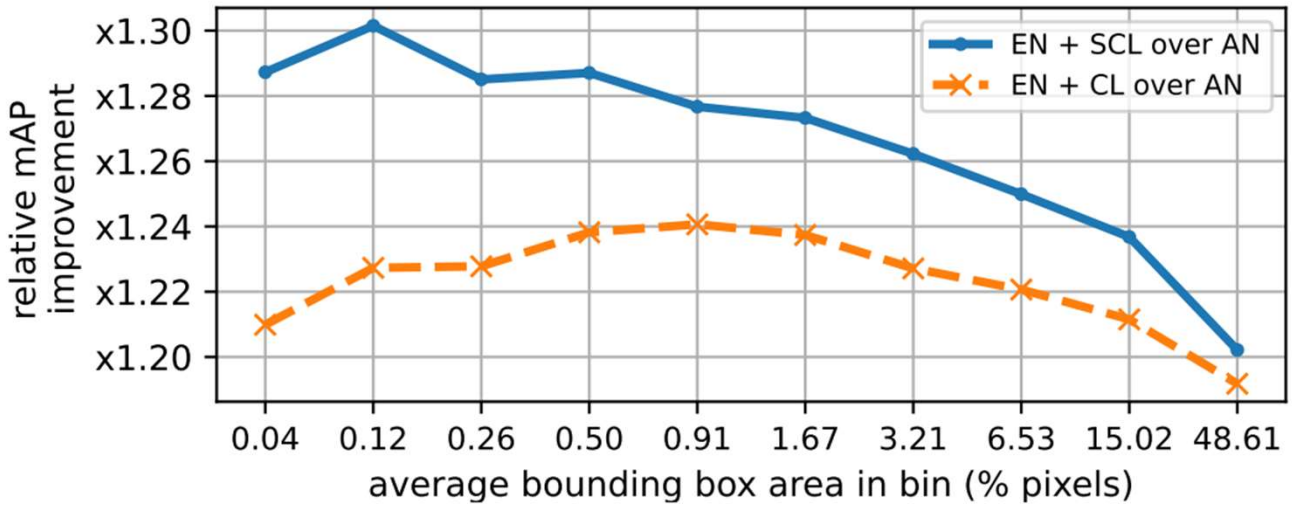


Figure D.1. Relative improvement per object size.

split the positive annotation into equal size, compute mAP use positive labels within bin, and negative over whole val.

Analysis over object sizes

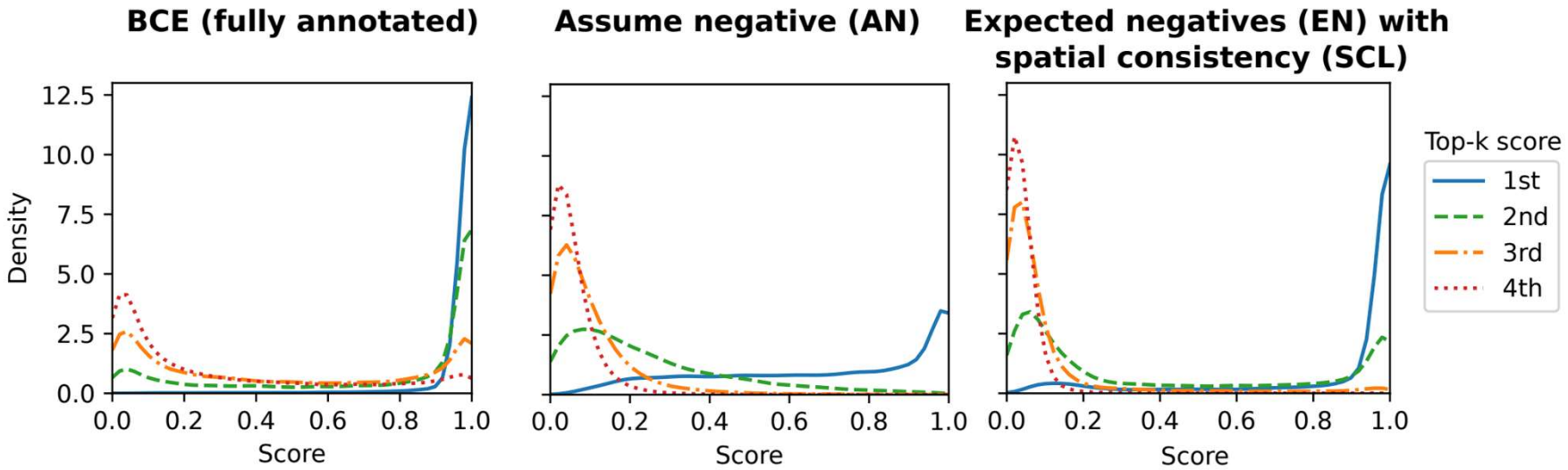


Figure F.1. Score distribution over all MS-COCO validation images, for 1st, 2nd, 3rd and 4th highest predicted scores per image. The BCE method is a fully annotated baseline. Training with AN and a single-positive label leads to a bias towards single positive predictions. With EN and SCL, the network more confidently predicts multiple positives.

An : low scores distribution,

EN+SCL : reduce the amount of false negative label,

Experiments

	top-1 IN-val	top-1 ReaL	mAP ReaL				
			k = all	k = 1	k = 2	k = 3	k $\geq$ 4
Num. samples	50,000	46,837	46,837	39,394	5,408	1,319	716
ResNet-50	76.1	83.0	66.3	70.6	53.0	36.1	**22.5**
ResNet-50 + AN	76.9	83.1	81.4	88.0	60.0	36.8	21.8
ResNet-50 + EN with CL	**77.1**	83.4	81.7	88.4	60.5	36.6	21.7
ResNet-50 + EN with SCL	**77.1**	**83.9**	**82.3**	**88.5**	**61.9**	**38.1**	**22.5**

Table 3. We finetune ResNet-50 with AN, consistency loss (CL) or spatial consistency loss (SCL). We report top-1 validation accuracy on ImageNet-val (single-label) and on ReaL (multi-label); as well as mean average precision (mAP) on ReaL. mAP is reported on all images (k = all), or on subsets of images with $k = 1, 2, 3, 4+$ annotated labels.

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

THANKS