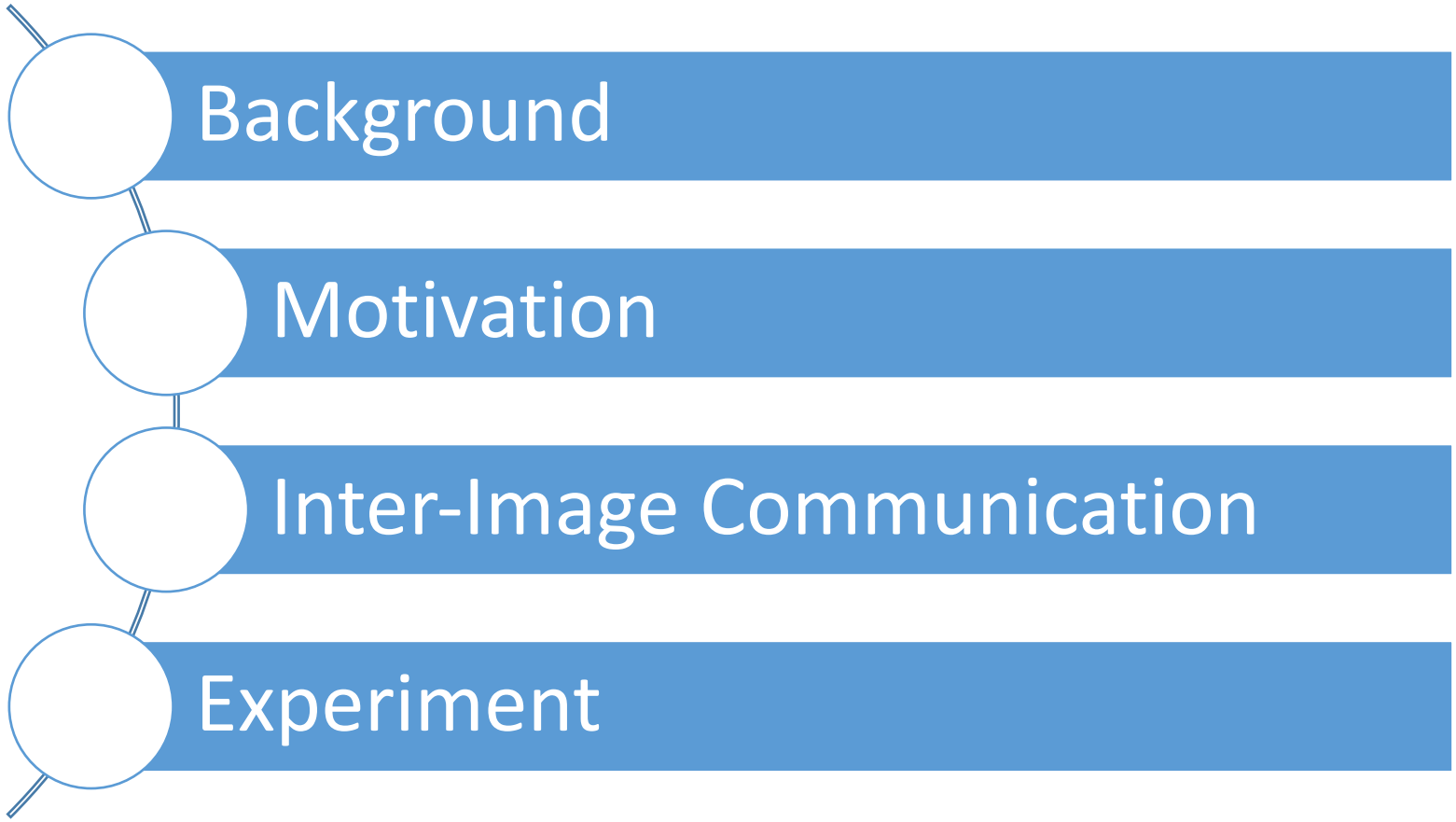
Inter-Image Communication for Weakly Supervised Localization

Xiaolin Zhang, Yunchao Wei, Yi Yang

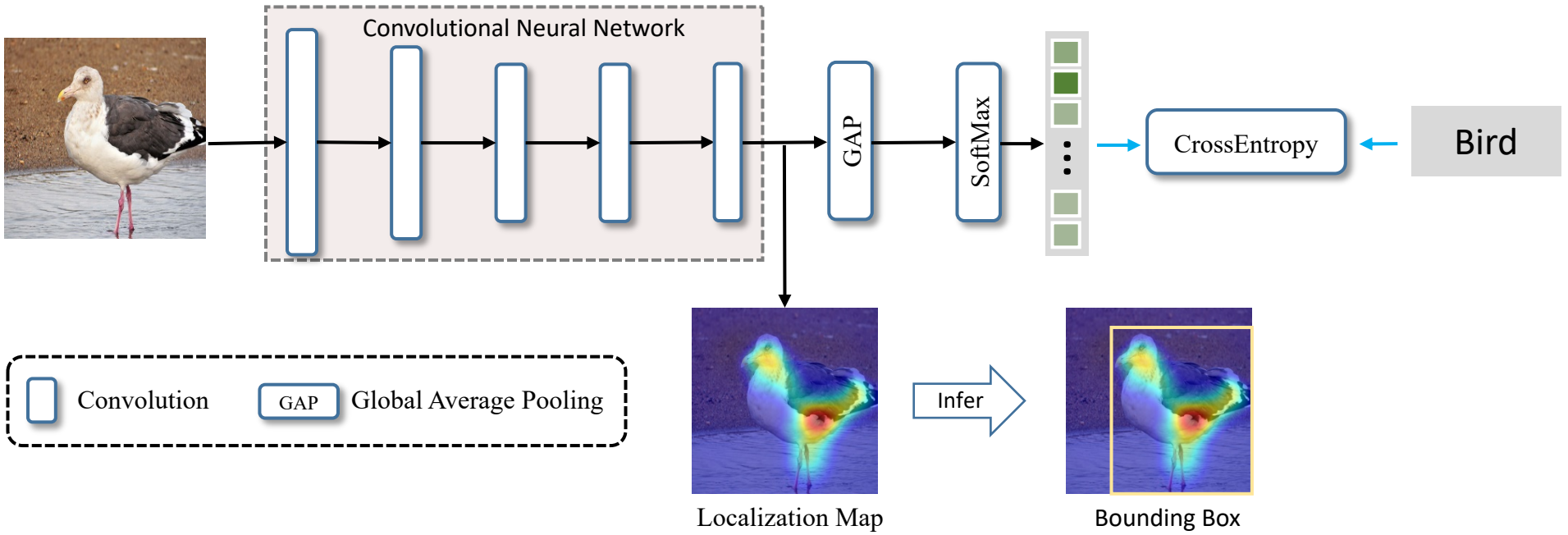
ReLER, AAII, University of Technology Sydney

http://arxiv.org/abs/2008.05096

Content

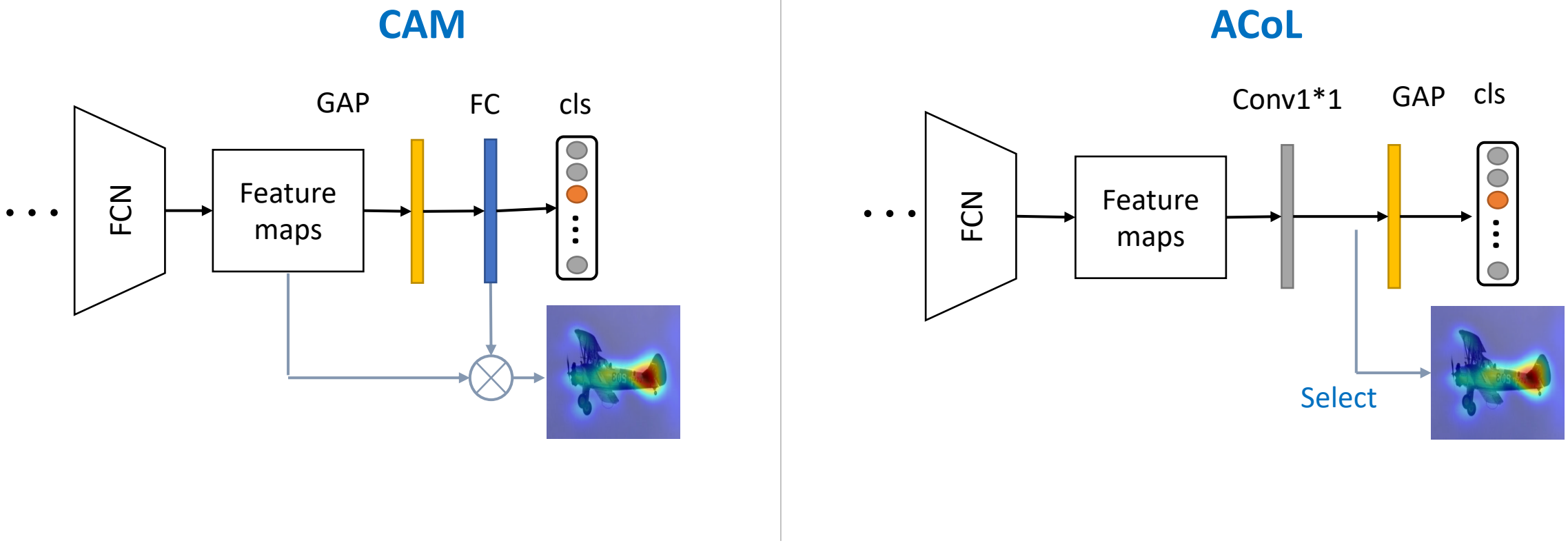

Background
Motivation
Inter-Image Communication
Experiment

Weakly Supervised Localization



- Given an image-level label, WSL is to find the location of the target object.
- In practice, classification networks are trained to extract localization maps. Bounding boxes are inferred offline.
- **The fundamental target of WSL is to improve the localization maps.**

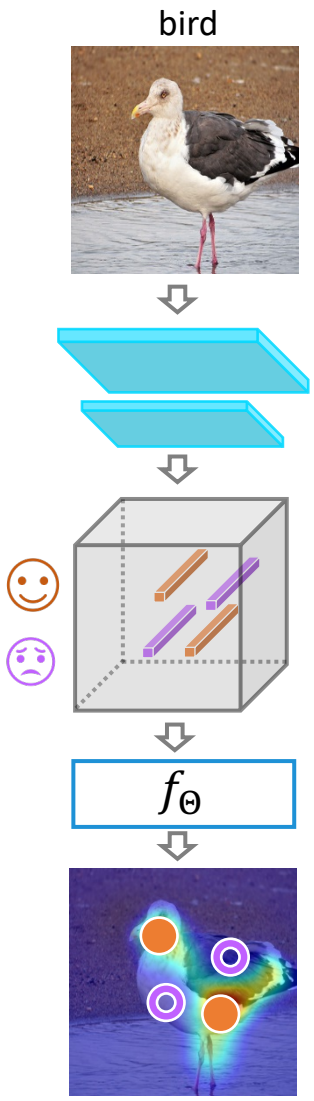
Methods for extracting localization maps



Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. IEEE CVPR (2016)
Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.: Adversarial complementary learning for weakly supervised object localization. In: IEEE CVPR (2018)

Motivation

Current



Observations and Facts

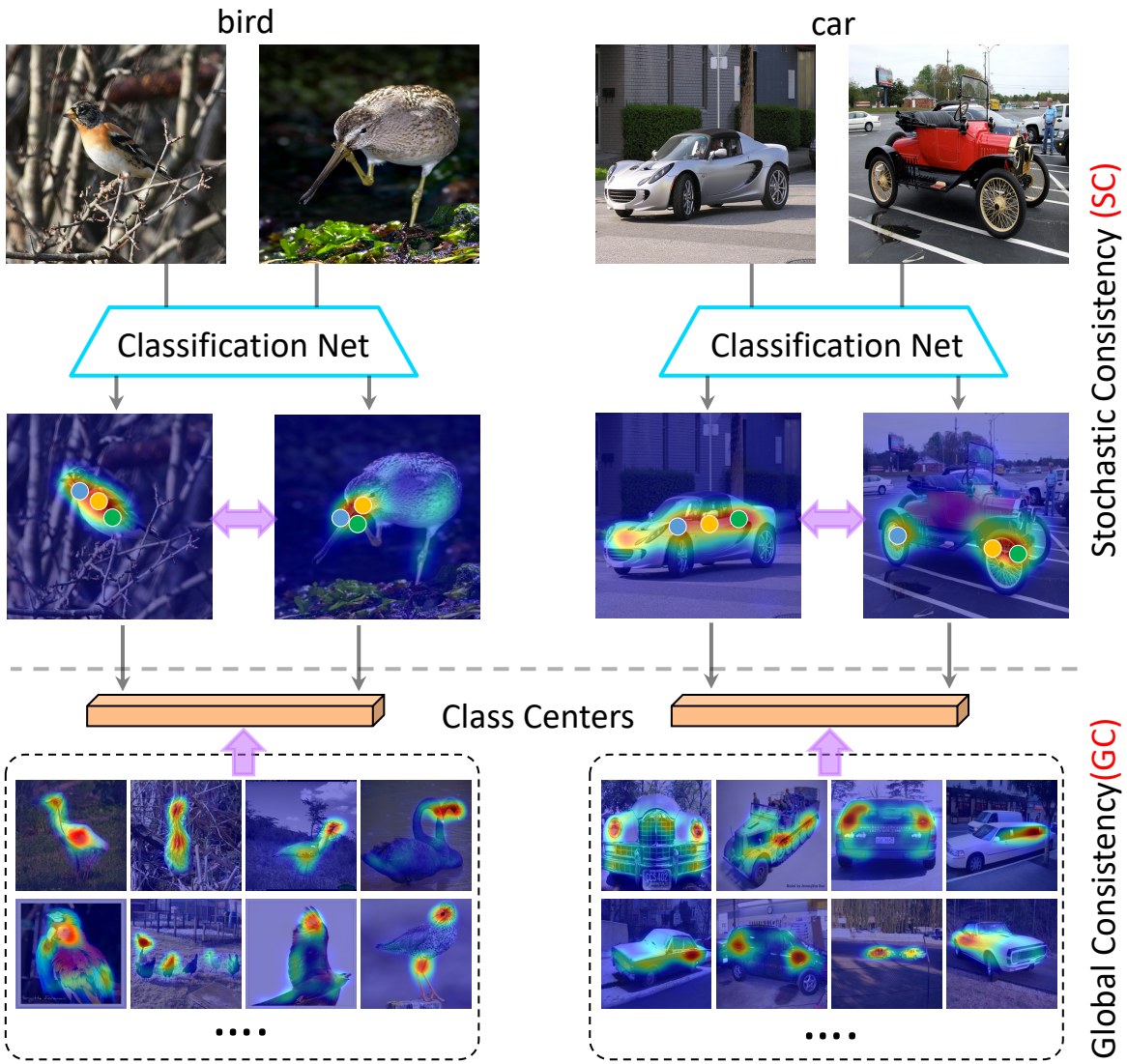
- Some regions belonging to the object are not highlighted
- In general, vectors in orange/purple are responsible to the scores in orange/purple.
- Images are processed independently

Intuition of I2C

- Push the vectors belonging to the same object close
- Push the object vectors across images in a same category close (Inter-image communication)

Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: ECCV. Springer (2018)

Our

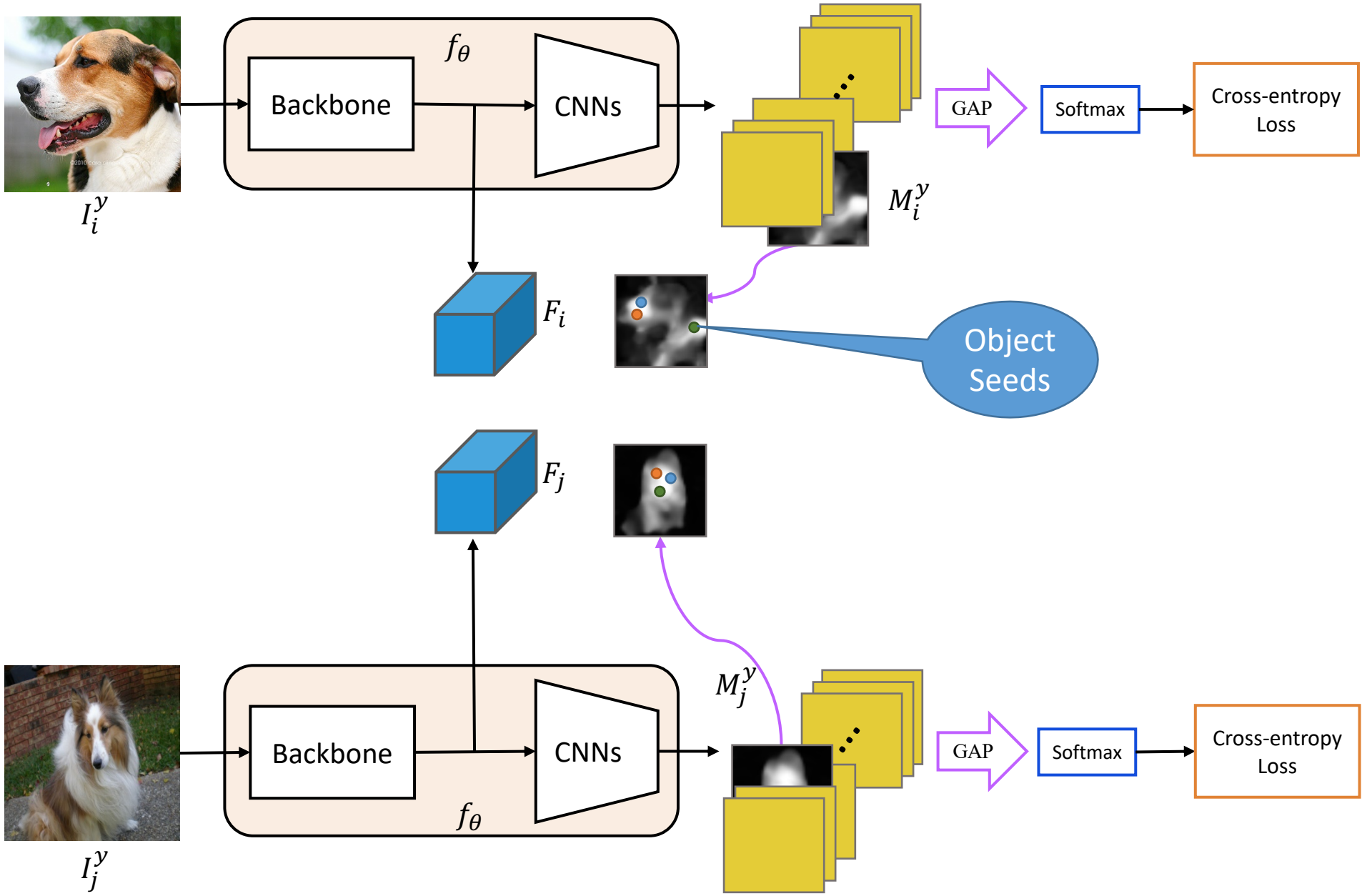


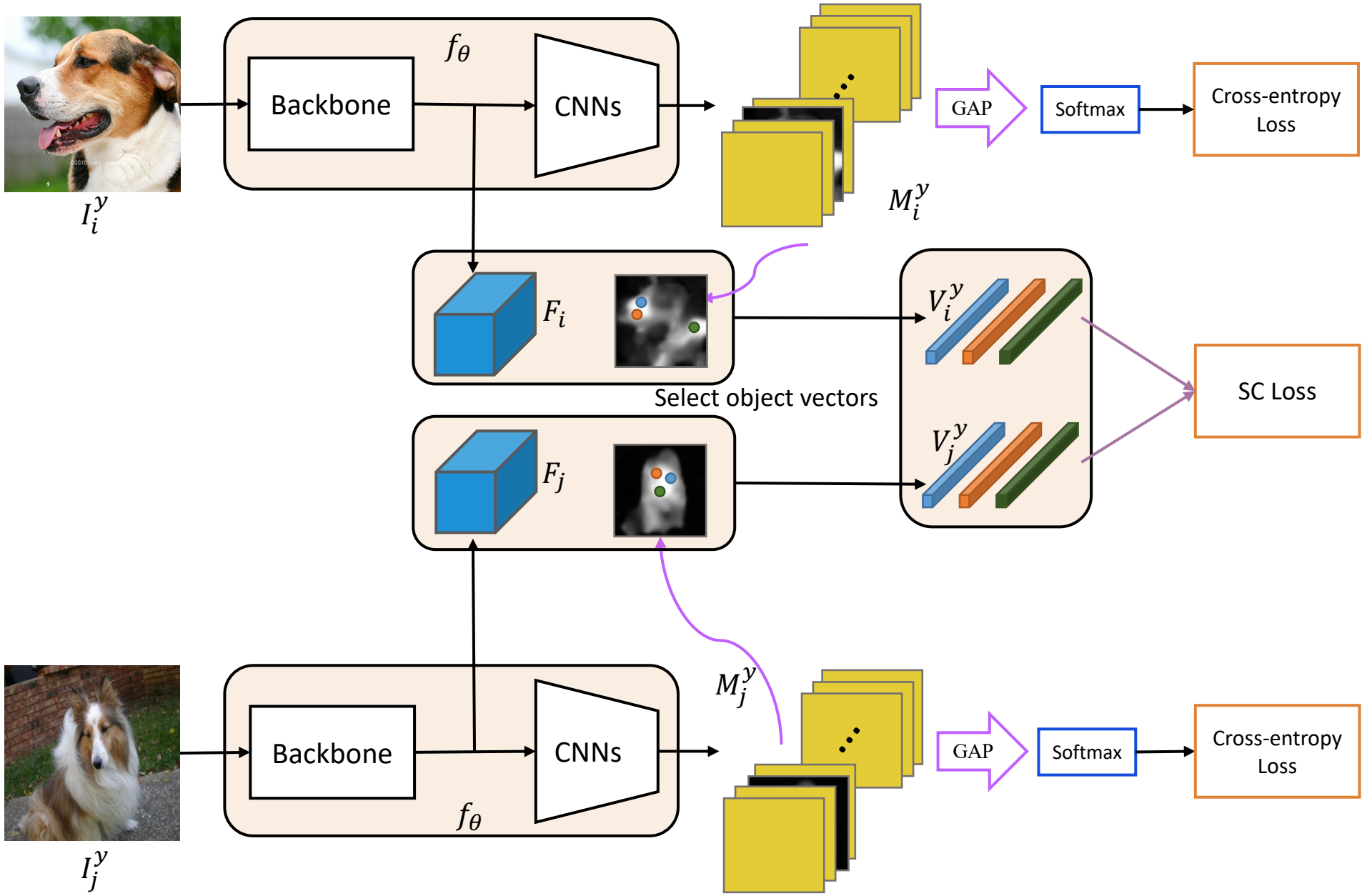
Stochastic Consistency (SC)

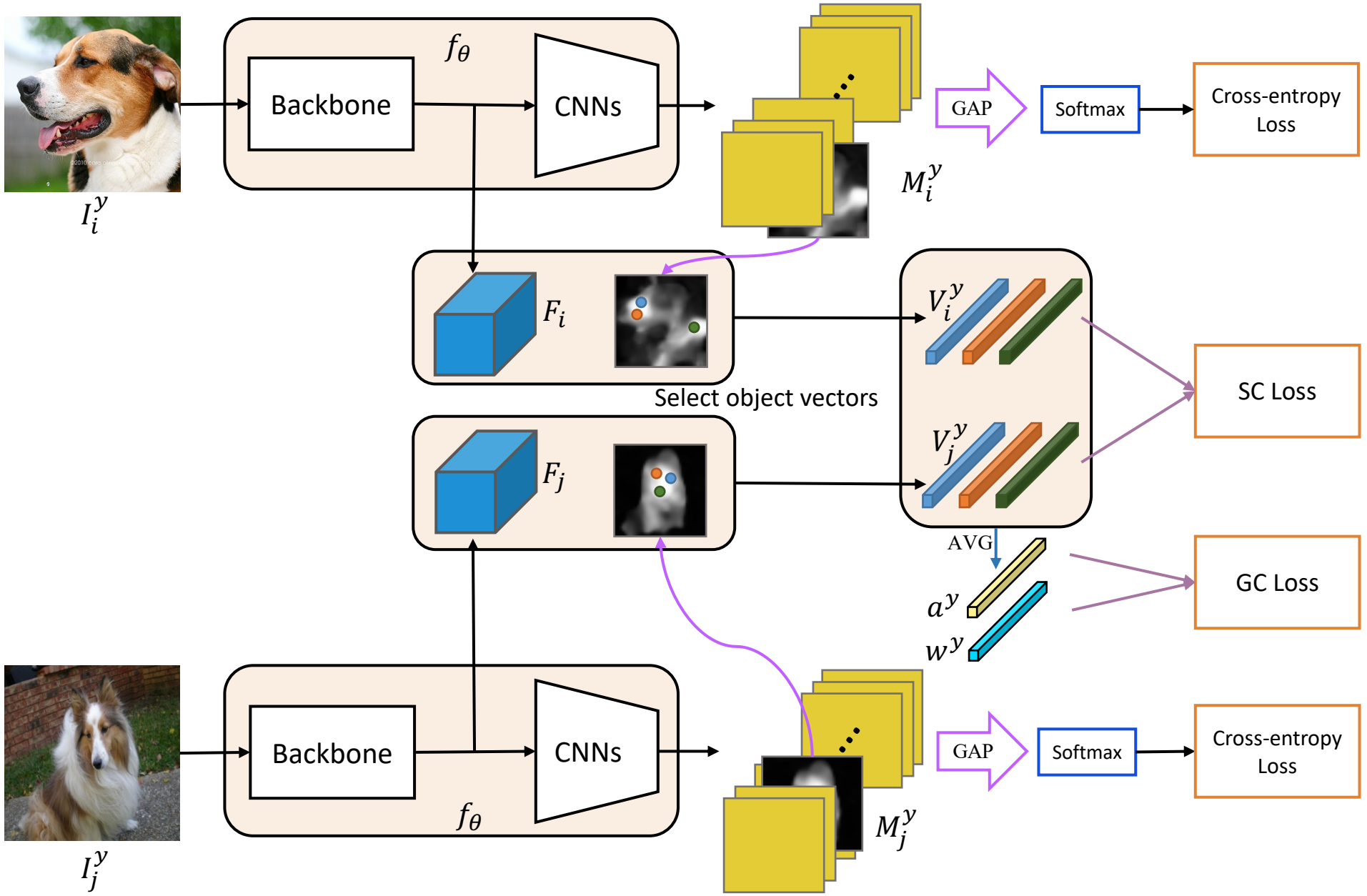
- Push feature vectors of the objects from two images close
- Operate within a minibatch

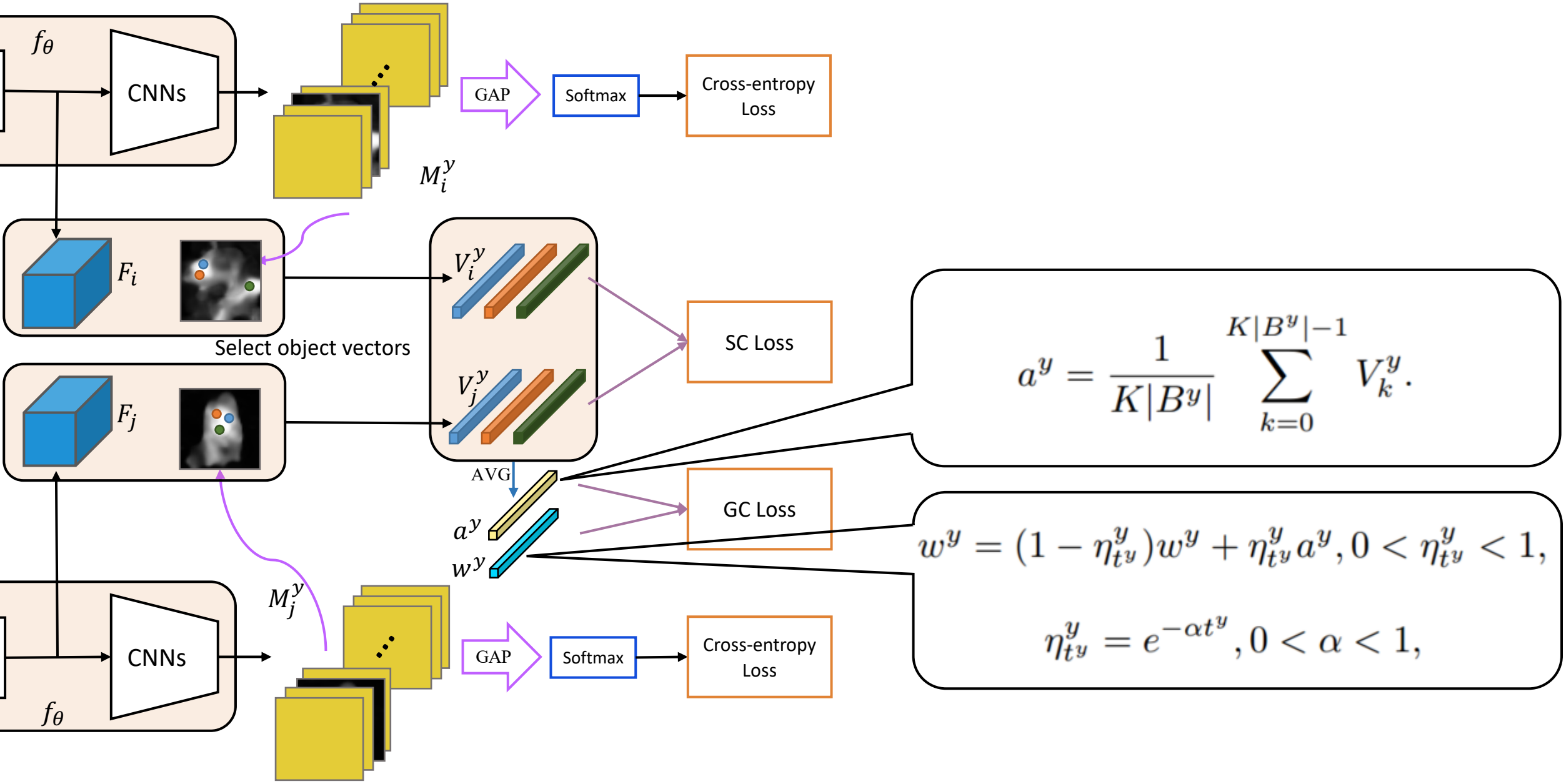
Global Consistency (GC)

- Push feature vectors approach their class centers
- One center vector for each category
- Operate across minibatches


I_i^y
Backbone
f_θ
CNNs
M_i^y
GAP
Softmax
Cross-entropy Loss
F_i
Object Seeds
F_j
M_j^y
Backbone
CNNs
f_θ
GAP
Softmax
Cross-entropy Loss
I_j^y


I_i^y
Backbone
f_θ
CNNs
M_i^y
GAP
Softmax
Cross-entropy Loss
F_i
Select object vectors
V_i^y
V_j^y
SC Loss
F_j
I_j^y
Backbone
CNNs
f_θ
M_j^y
GAP
Softmax
Cross-entropy Loss


I_i^y
Backbone
f_θ
CNNs
M_i^y
GAP
Softmax
Cross-entropy Loss
F_i
Select object vectors
V_i^y
V_j^y
SC Loss
F_j
AVG
a^y
w^y
GC Loss
I_j^y
Backbone
CNNs
f_θ
M_j^y
GAP
Softmax
Cross-entropy Loss


f_θ
CNNs
GAP
Softmax
Cross-entropy Loss
M_i^y
F_i
Select object vectors
V_i^y
V_j^y
SC Loss
F_j
AVG
a^y
w^y
GC Loss
M_j^y
CNNs
f_θ
GAP
Softmax
Cross-entropy Loss
$a^y = \frac{1}{K|B^y|} \sum_{k=0}^{K|B^y|-1} V_k^y.$
$w^y = (1 - \eta_{t^y}^y) w^y + \eta_{t^y}^y a^y, 0 < \eta_{t^y}^y < 1,$
$\eta_{t^y}^y = e^{-\alpha t^y}, 0 < \alpha < 1,$

Object Seed

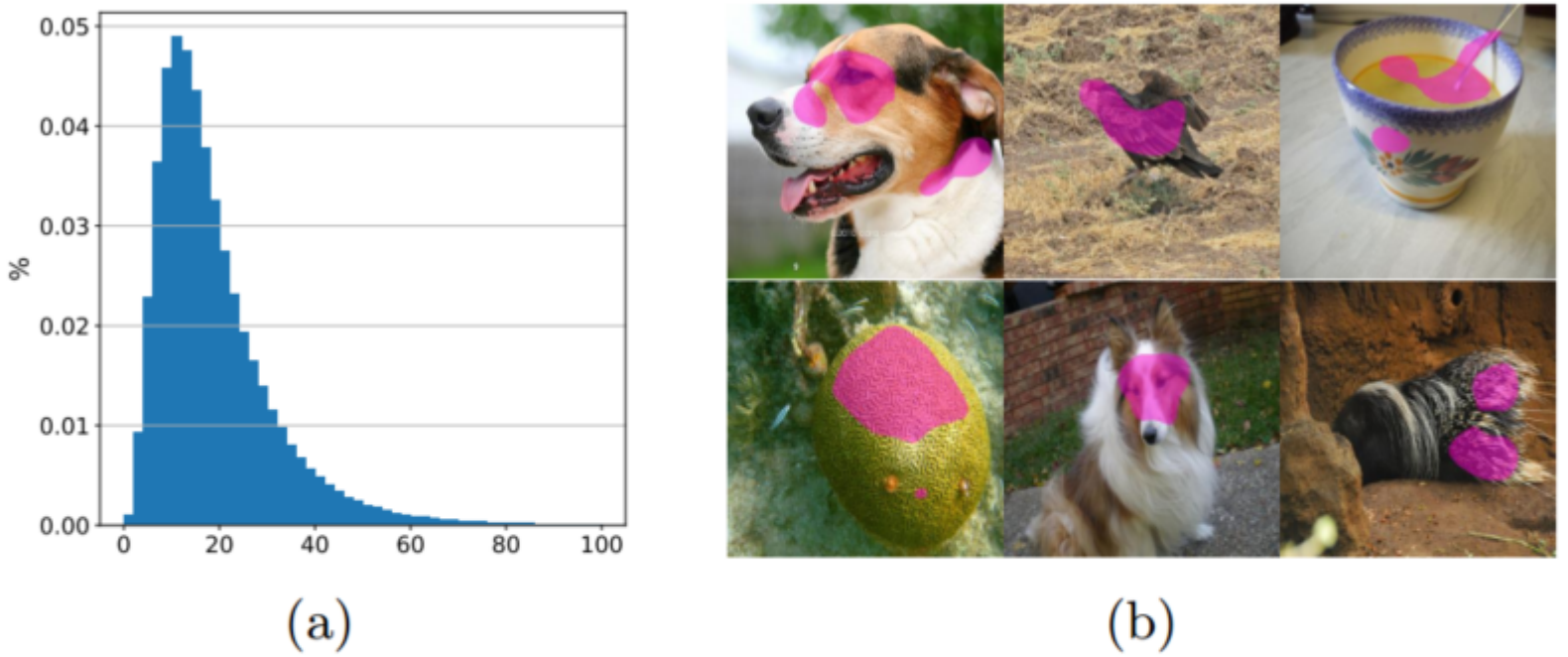
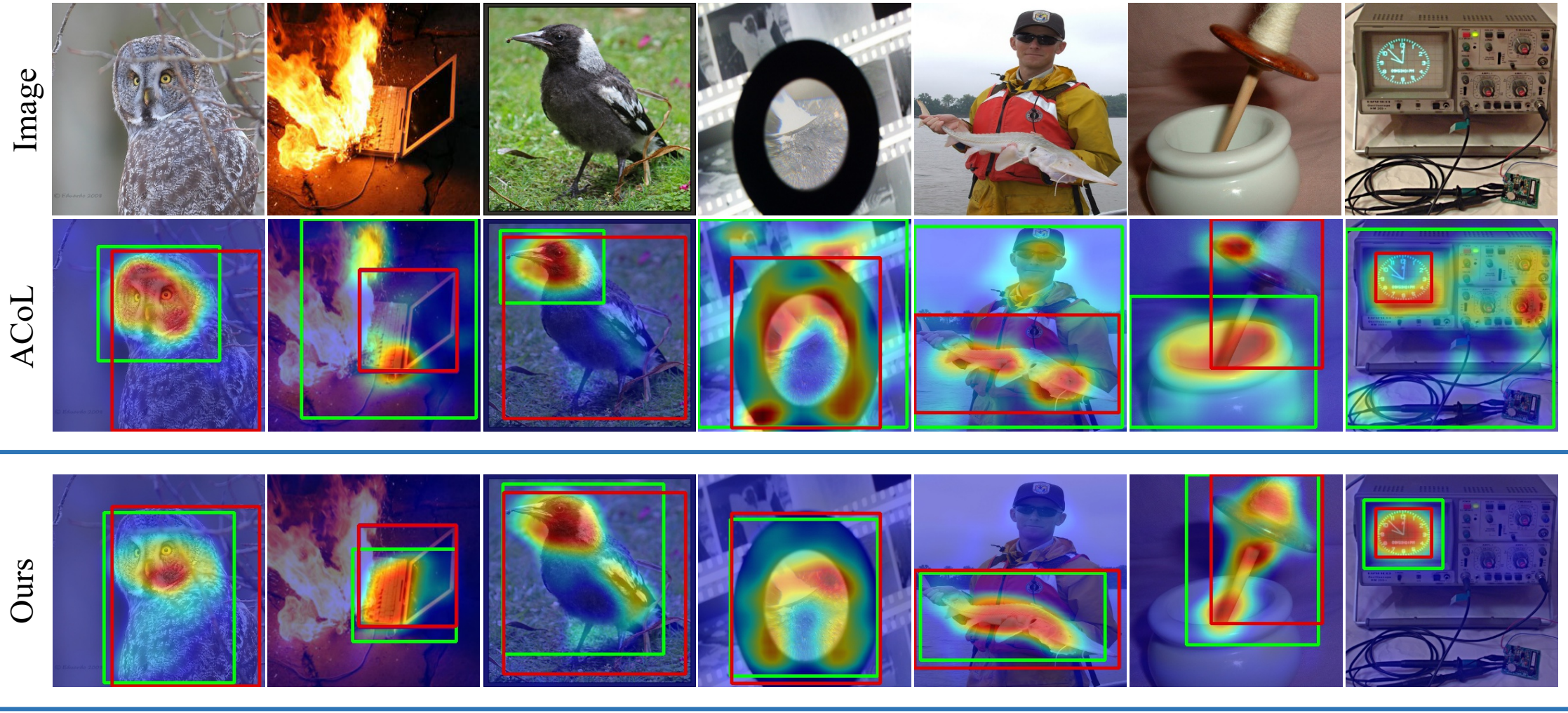


Fig. 4: *(a)*: histogram of the number of object pixels with the threshold of 0.7 on the ILSVRC training set. *(b)*: identified object regions (in magenta) according to the localization maps

Visualization Comparison



Predicted boxes are in green, while Ground-truth boxes are in red.

Comparison of Localization Error on ILSVRC

Methods	Backbone	Localization Error			Classification Error	
		Top-1	Top-5	Gt-known	Top-1	Top-5
CAM [46]	AlexNet [22]	67.19	52.16	45.01	42.6	19.5
CAM [46]	GoogLeNet [32]	56.40	43.00	41.34	31.9	11.3
HaS-32 [31]	GoogLeNet [32]	54.53	-	39.43	32.5	-
ACoL [42]	GoogLeNet [32]	53.28	42.58	-	29.0	11.8
DANet [40]	GoogLeNet [30]	52.47	41.72	-	27.5	8.6
Backprop [29]	VGG16 [30]	61.12	51.46	-	-	
CAM [46]	VGG16 [30]	57.20	45.14	-	31.2	11.4
CutMix [41]	VGG16 [30]	56.55	-	-	-	-
ADL [8]	VGG16 [30]	55.08	-	-	32.2	-
ACoL [42]	VGG16 [30]	54.17	**40.57**	37.04	32.5	12.0
I^2C-Ours	VGG16 [30]	**52.59**	41.49	**36.10**	30.6	10.7
CAM [46]	ResNet50-SE [15,17]	53.81	-	-	23.44	-
CutMix [41]	ResNet50 [15]	52.75	-	-	21.4	5.92
ADL [8]	ResNet50-SE [15,17]	51.47	-	-	24.15	-
I^2C-Ours	ResNet50 [15]	**45.17**	**35.40**	**31.50**	23.3	6.9
CAM [46]	InceptionV3 [33]	53.71	41.81	37.32	26.7	8.2
SPG [43]	InceptionV3 [33]	51.40	40.00	35.31	30.3	9.9
ADL [8]	InceptionV3 [33]	51.29	-	-	27.2	-
I^2C-Ours	InceptionV3 [33]	**46.89**	**35.87**	**31.50**	26.7	8.4

Table 1: Comparison of the localization error rate on ILSVRC validation set. Classification error rates is also presented for reference.

Comparison of Localization Error on CUB

Methods	Top-1	Top-5
CAM [46]	56.33	46.47
ACoL [42]	54.08	43.49
SPG [43]	53.36	42.28
DANet [40]	47.48	38.04
ADL [8]	46.96	-
I^2C-Ours	**44.01**	**31.66**

Table 2: Localization error on the CUB-200-2011 test set. I^2C significantly surpasses all the baselines.

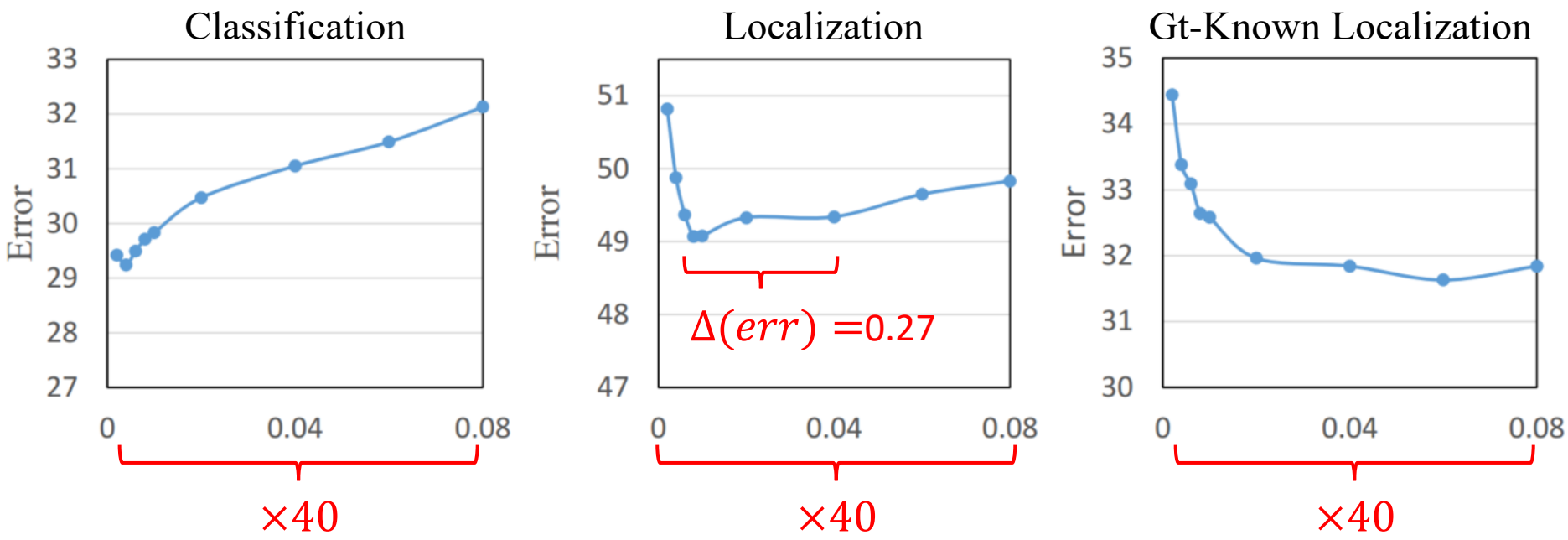
Are SC and GC really effective?

Methods	Plain	SC	SC + GC
Loc.	53.71*	49.07	**48.08**
Gt-known Loc.	37.32	31.63	**31.04**

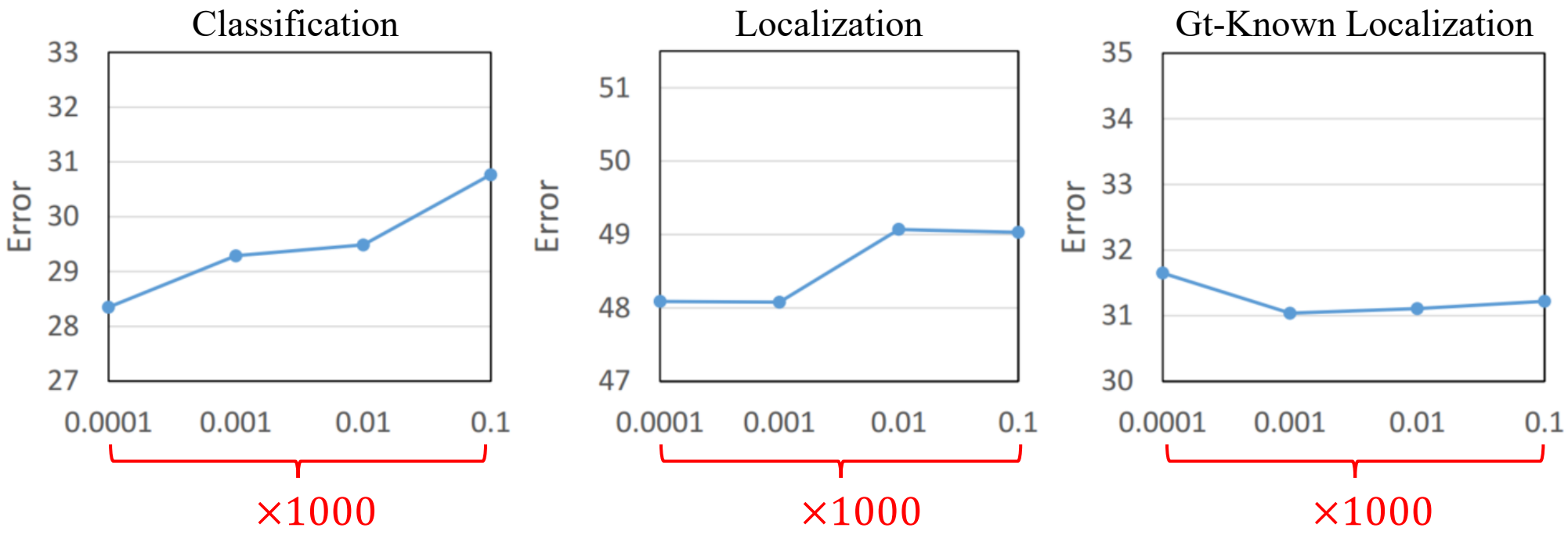
Table 4: Localization error on ILSVRC validation set using different configurations of the proposed constraints. ($*$ indicates the numbers obtained with the classification results using the ten-crop operation.)

How do the hyper-variables affect the localization performance?

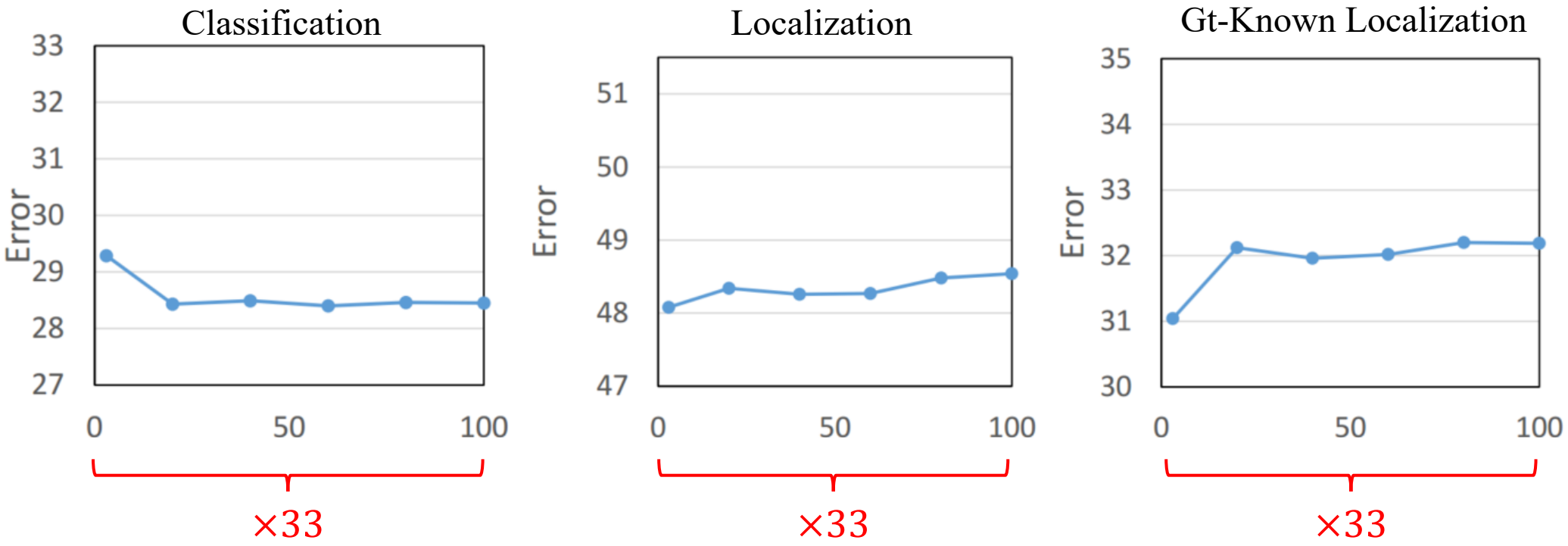
λ_1


Classification
Error
33
32
31
30
29
28
27
0
0.04
0.08
×40
Localization
Error
51
50
49
48
47
Δ(err) =0.27
0
0.04
0.08
×40
Gt-Known Localization
Error
35
34
33
32
31
30
0
0.04
0.08
×40

λ_2


Classification
Error
33
32
31
30
29
28
27
0.0001
0.001
0.01
0.1
×1000
Localization
Error
51
50
49
48
47
0.0001
0.001
0.01
0.1
×1000
Gt-Known Localization
Error
35
34
33
32
31
30
0.0001
0.001
0.01
0.1
×1000

K


Classification
Error
33
32
31
30
29
28
27
0
50
100
×33
Localization
Error
51
50
49
48
47
0
50
100
×33
Gt-Known Localization
Error
35
34
33
32
31
30
0
50
100
×33

Thank You