Fine-grained Image Retrieval

by Jiawei

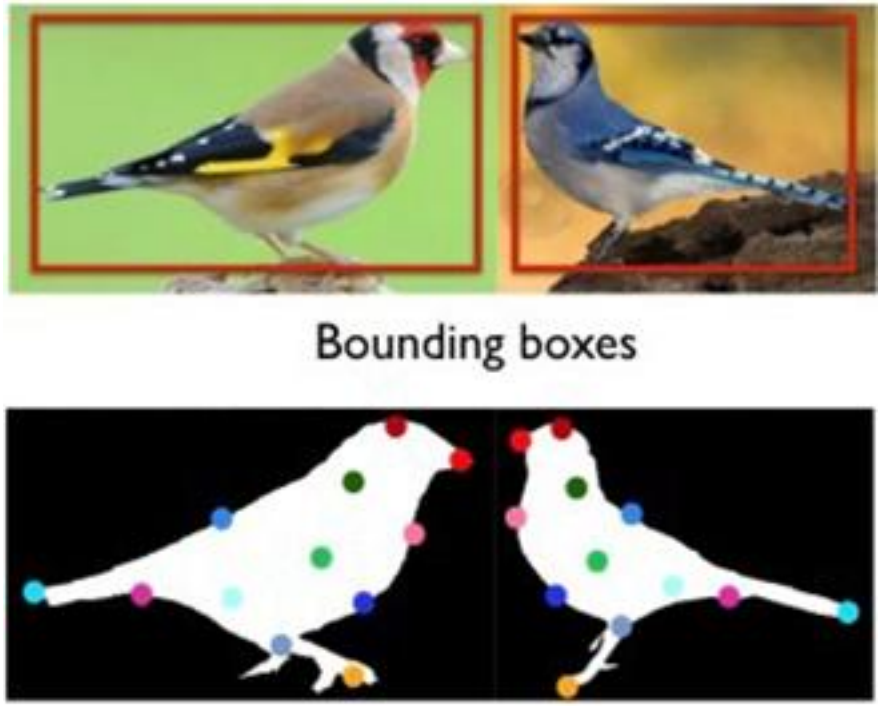
Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval

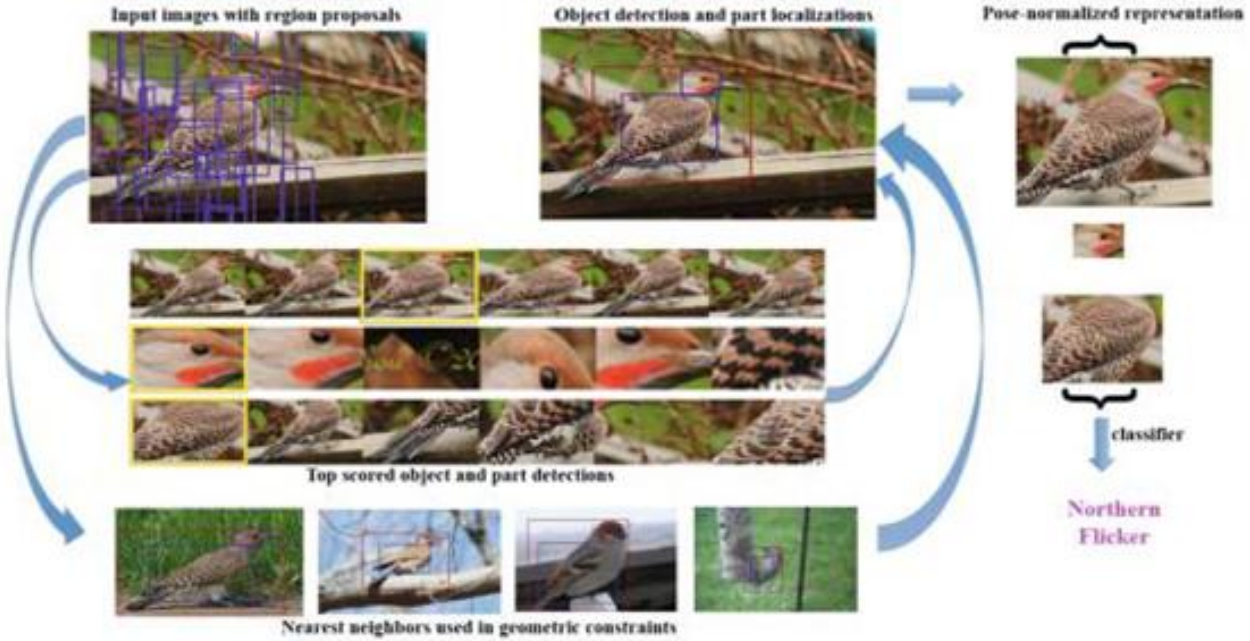
Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, Member, IEEE, Zhi-Hua Zhou, Fellow, IEEE

About Fine-Grained Image

- 细粒度图像分类
- 细粒度图像检索

细粒度图像分类


Bounding boxes


Input images with region proposals
Object detection and part localizations
Pose-normalized representation
Top scored object and part detections
Nearest neighbors used in geometric constraints
classifier
Northern Flicker

图像检索

(b) General image retrieval. Two examples from the *Oxford Building* [12] dataset.

Fine-grained *retrieval*

(a) Fine-grained image retrieval. Two examples ("Mallard" and "Rolls-Royce Phantom Sedan 2012") from the *CUB200-2011* [10] and *Cars* [11] datasets, respectively.

Abstract

- convolutional neural network models pre- trained for the ImageNet classification task
- propose the Selective Convolutional Descriptor Aggre- gation (SCDA) method

SCDA

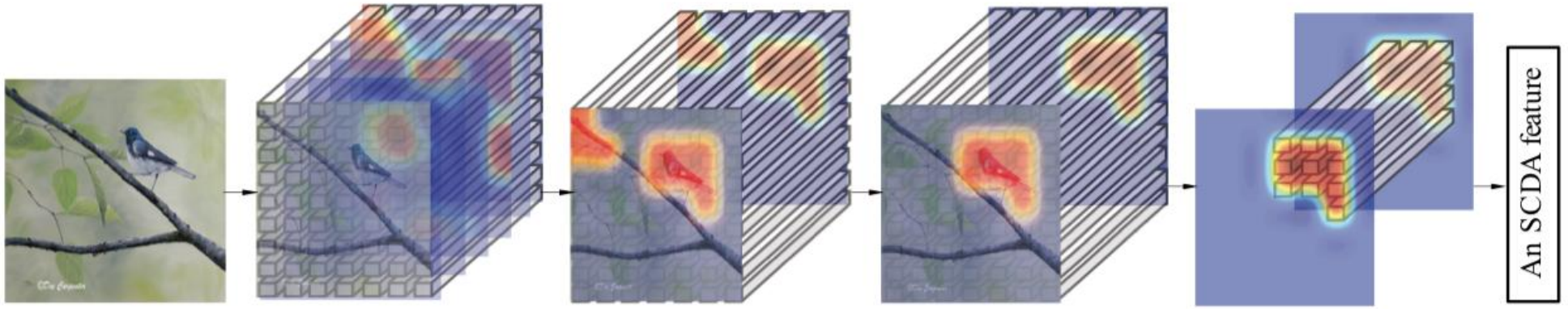
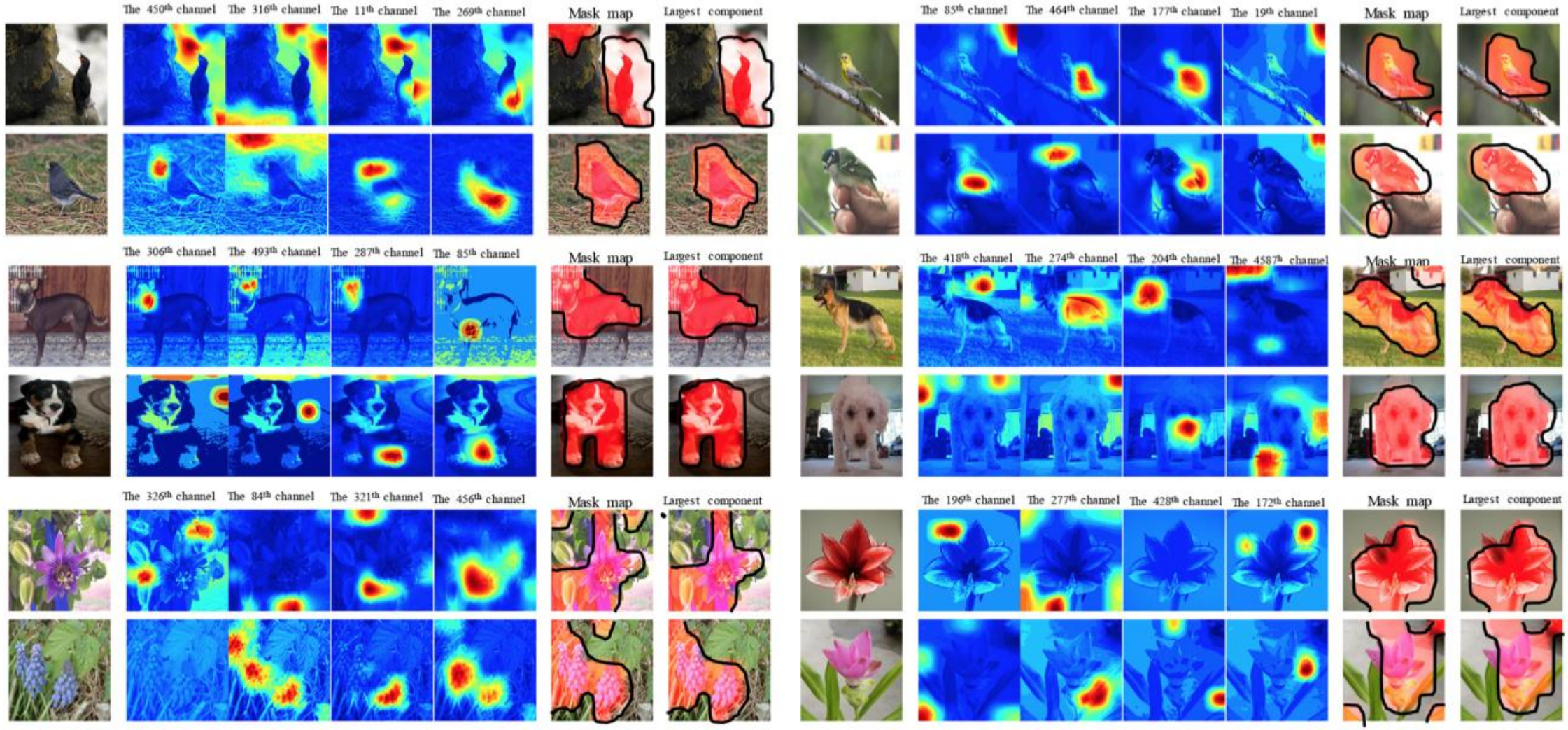


Figure 2. Pipeline of the proposed SCDA method. An input image with arbitrary resolution is fed into a pre-trained CNN model, and extracted as an order-3 convolution activation tensor. Based on the activation tensor, SCDA firstly selects the deep descriptors by locating the main object in fine-grained images unsupervisedly. Then, it pools the selected deep descriptors into the SCDA feature as the whole image representation. In the figure, (b)-(e) show the process of selecting useful deep convolutional descriptors, and the details can be found in Sec. III-B1. (This figure is best viewed in color.)

- using only the pre-trained model
- Each concept is represented by a pattern of activity distributed over many neurons, and each neuron participates in the representation of many concepts
- Fig. 3 conveys that not all deep descriptors are useful, and one single channel contains at best weak semantic information due to the distributed nature of this representation.


The 450th channel
The 316th channel
The 11th channel
The 269th channel
Mask map
Largest component
The 85th channel
The 464th channel
The 177th channel
The 19th channel
Mask map
Largest component
The 306th channel
The 493th channel
The 287th channel
The 85th channel
Mask map
Largest component
The 418th channel
The 274th channel
The 204th channel
The 4587th channel
Mask map
Largest component
The 326th channel
The 84th channel
The 321th channel
The 456th channel
Mask map
Largest component
The 196th channel
The 277th channel
The 428th channel
The 172th channel
Mask map
Largest component

- Consequently, we calculate the mean value $\bar{a}$ of all the positions in A as the threshold to decide which positions localize objects

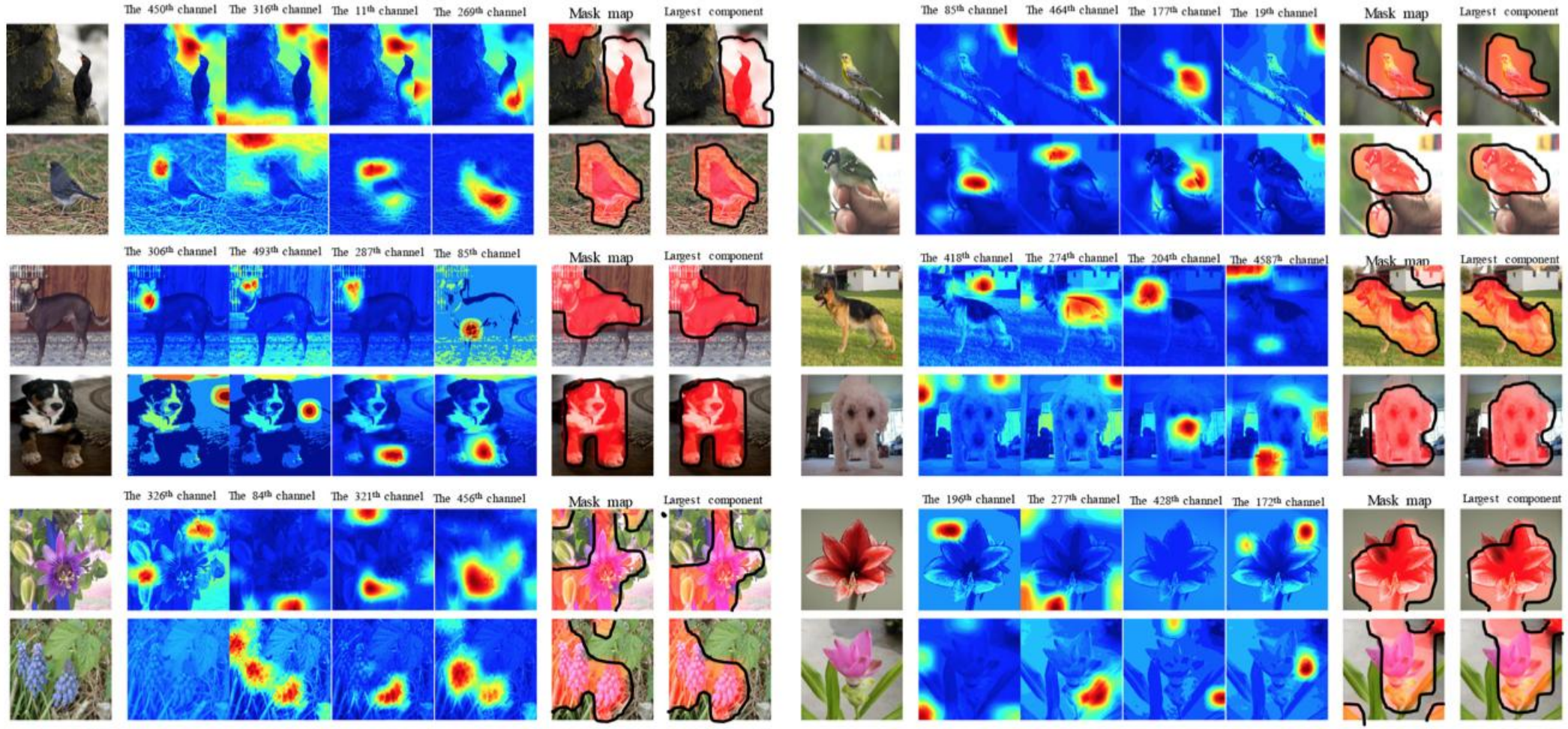
$$M_{i,j} = \begin{cases} 1 & \text{if } A_{i,j} > \bar{a} \\ 0 & \text{otherwise} \end{cases},$$

- we employ Algorithm 1 to collect the largest connected component of M

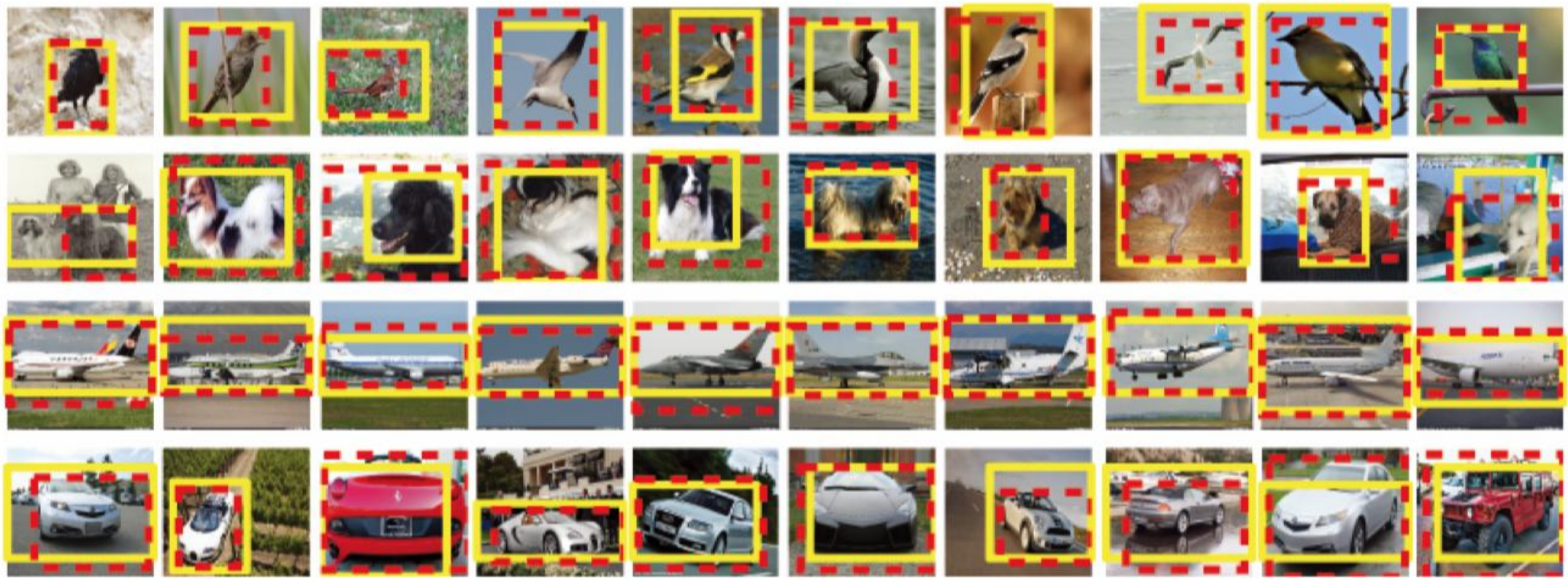
Algorithm 1 Finding connected components in binary images

Require: A binary image I;

1: Select one pixel p as the starting point;
2: **while** True **do**
3: Use a flood-fill algorithm to label all the pixels in the connected component containing p;
4: **if** All the pixels are labeled **then**
5: Break;
6: **end if**
7: Search for the next unlabeled pixel as p;
8: **end while**
9: **return** Connectivity of the connected components, and their corresponding size (pixel numbers).


The 450th channel
The 316th channel
The 11th channel
The 269th channel
Mask map
Largest component
The 85th channel
The 464th channel
The 177th channel
The 19th channel
Mask map
Largest component
The 306th channel
The 493th channel
The 287th channel
The 85th channel
Mask map
Largest component
The 418th channel
The 274th channel
The 204th channel
The 4587th channel
Mask map
Largest component
The 326th channel
The 84th channel
The 321th channel
The 456th channel
Mask map
Largest component
The 196th channel
The 277th channel
The 428th channel
The 172th channel
Mask map
Largest component

- Qualitative Evaluation
- Because four fine-grained datasets (i.e., CUB200- 2011, Stanford Dogs, Aircrafts and Cars) supply the ground- truth bounding box for each image, it is desirable to evaluate the proposed method for object localization. However, as seen in Fig. 3, the detected regions are irregularly shaped.닸So, the minimum rectangle bounding boxes which contain the detected regions are returned as our object localization predictions.

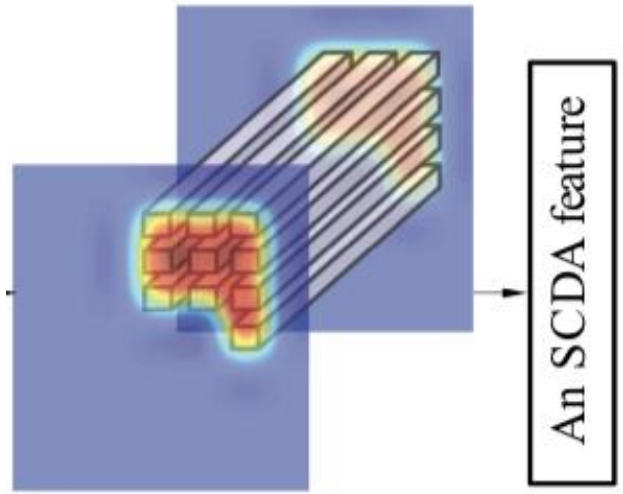


Quantitative Evaluation

- The reported metrics are the percentage of whole-object boxes that are correctly localized with a >50% IOU with the ground-truth bounding boxes.

Dataset	Method	Train phase		Test phase		Head	Torso	Whole-object
		BBox	Parts	BBox	Parts			
CUB200-2011	Strong DPM [38]	✓	✓	✓		43.49	75.15	–
	Part-based R-CNN with BBox [4]	✓	✓	✓		68.19	79.82	–
	Deep LAC [5]	✓	✓	✓		74.00	96.00	–
	Part-based R-CNN [4]	✓	✓			61.42	70.68	–
	Unsupervised object discovery [39]					–	–	69.37
	Ours					–	–	76.79
Stanford Dogs	Unsupervised object discovery [39]					–	–	36.23
	Ours					–	–	78.86
Aircrafts	Unsupervised object discovery [39]					–	–	42.11
	Ours					–	–	94.91
Cars	Unsupervised object discovery [39]					–	–	93.05
	Ours					–	–	90.96

Aggregating Convolutional Descriptors



(e) Selected descriptors (f) Descriptors aggregation

Table II

COMPARISON OF DIFFERENT ENCODING OR POOLING APPROACHES FOR FGIR. THE BEST RESULT OF EACH COLUMN IS MARKED IN BOLD.

Approach	Dimension	*CUB200-2011*		*Stanford Dogs*	
		top1	top5	top1	top5
VLAD (k=2)	1,024	55.92	62.51	69.28	74.43
VLAD (k=128)	6,5536	55.66	62.40	68.47	75.01
Fisher Vector (k=2)	2,048	52.04	59.19	68.37	73.74
Fisher Vector (k=128)	131,072	45.44	53.10	61.40	67.63
avgPool	512	56.42	63.14	73.76	78.47
maxPool	512	58.35	64.18	70.37	75.59
avg&maxPool	1,024	**59.72**	**65.79**	**74.86**	**79.24**

Result

Table III
COMPARISON OF FINE-GRAINED IMAGE RETRIEVAL PERFORMANCE. THE BEST RESULT OF EACH COLUMN IS IN BOLD.

Method	Dimension	*CUB200-2011*		*Stanford Dogs*		*Oxford Flowers*		*Oxford Pets*		*Aircrafts*		*Cars*	
		top1	top5	top1	top5	top1	top5	top1	top5	top1	top5	top1	top5
SIFT_FV	32,768	5.25	8.07	12.58	16.38	30.02	36.19	17.50	24.97	30.69	37.44	19.30	24.11
SIFT_FV_gtBBox	32,768	9.98	14.29	15.86	21.15	–	–	–	–	38.70	46.87	34.47	40.34
fc_8_im	4,096	39.90	48.10	66.51	72.69	55.37	60.37	82.26	86.02	28.98	35.00	19.52	25.77
fc_8_gtBBox	4,096	47.55	55.34	70.41	76.61	–	–	–	–	34.80	41.25	30.02	37.45
fc_8_predBBox	4,096	45.24	53.05	68.78	74.09	57.16	62.24	85.55	88.47	30.42	36.50	22.27	29.24
$pool_5$	1,024	57.54	63.66	69.98	75.55	70.73	74.05	85.09	87.74	47.37	53.61	34.88	41.86
selectFV	2,048	52.04	59.19	68.37	73.74	70.47	73.60	85.04	87.09	48.69	54.68	35.32	41.60
selectVLAD	1,024	55.92	62.51	69.28	74.43	73.62	76.86	85.50	87.94	50.35	56.37	37.16	43.84
SPoC (w/o cen.)	256	34.79	42.54	48.80	55.95	71.36	74.55	60.86	67.78	37.47	43.73	29.86	36.23
SPoC (with cen.)	256	39.61	47.30	48.39	55.69	65.86	70.05	64.05	71.22	42.81	48.95	27.61	33.88
CroW	256	53.45	59.69	62.18	68.33	73.67	76.16	76.34	80.10	53.17	58.62	44.92	51.18
R-MAC	512	52.24	59.02	59.65	66.28	76.08	78.19	76.97	81.16	48.15	54.94	**46.54**	**52.98**
SCDA	1,024	59.72	65.79	74.86	79.24	75.13	77.70	87.63	89.26	53.26	58.64	38.24	45.16
$SCDA^+$	2,048	59.68	65.83	74.15	78.54	75.98	78.49	87.99	89.49	53.53	59.11	38.70	45.65
$SCDA\_flip^+$	4,096	**60.65**	**66.75**	**74.95**	**79.27**	**77.56**	**79.77**	**88.19**	**89.65**	**54.52**	**59.90**	40.12	46.73