Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval

Xiu-Shen Wei, Jian-Hao Luo and Jianxin Wu*

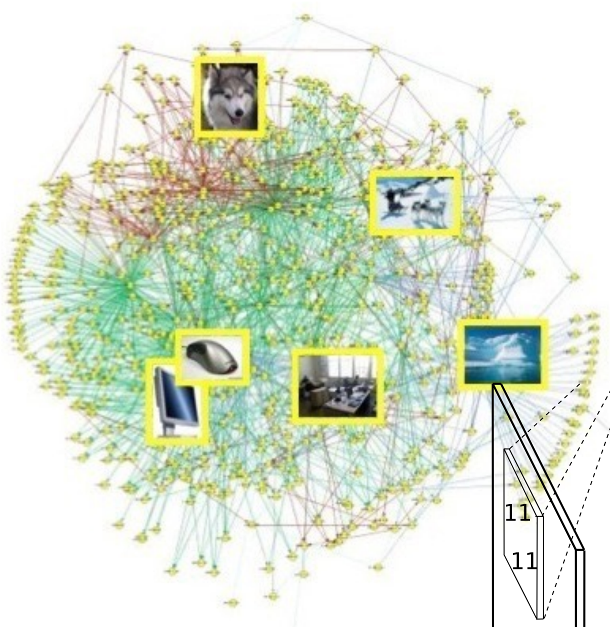
LAMDA Group
National Key Laboratory for Novel Software Technology
Nanjing University, China

Outline

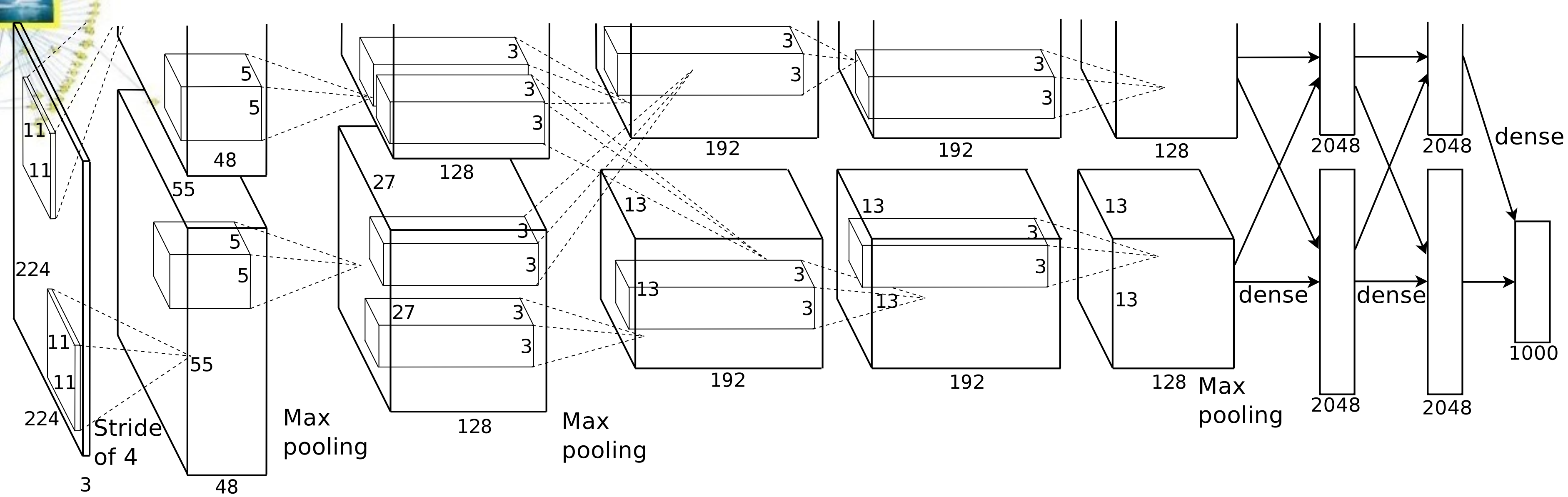
- Background
- Related works
- The proposed SCDA method
- Experiments
- Conclusions and future work

Background

Deep convolutional neural networks



IMAGENET



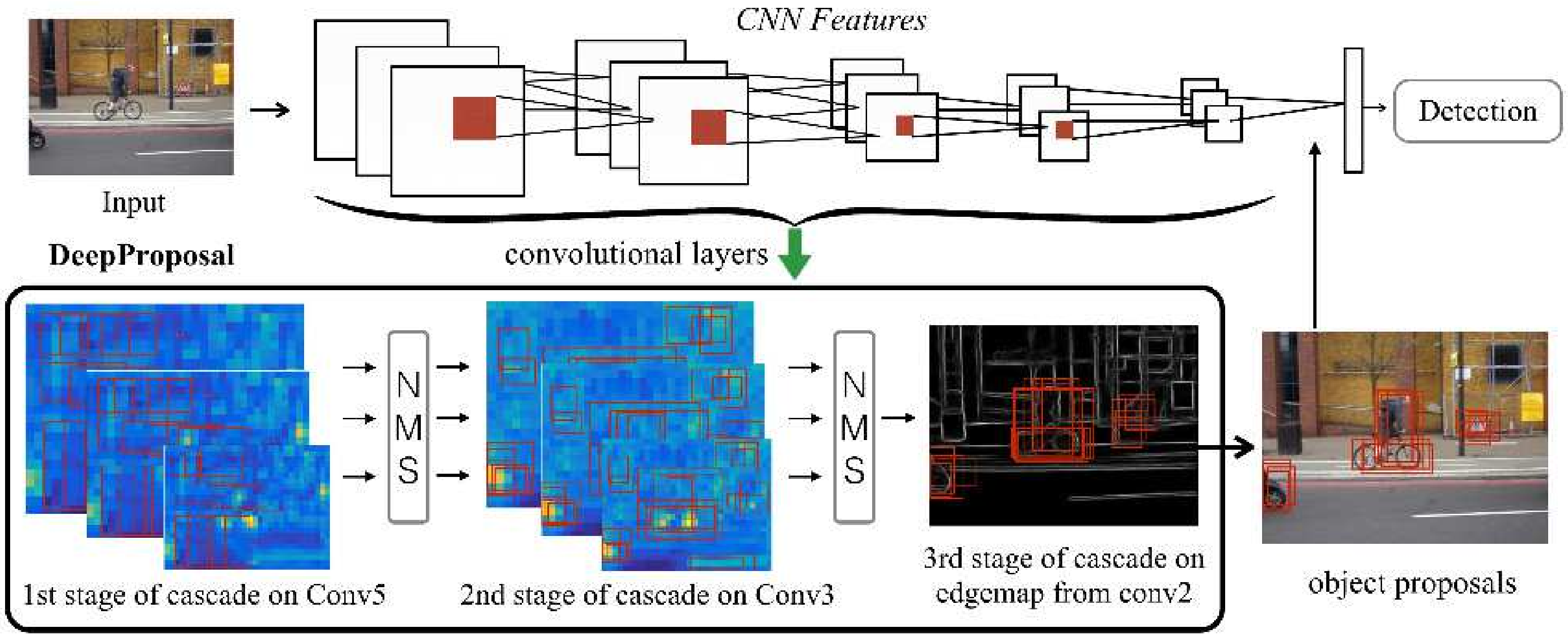
Background (con't)

Utilities of pre-trained deep models

[Cimpoi et al., CVPR'15] [Ghodrati et al., ICCV'15]

Background (con't)

Utilities of pre-trained deep models

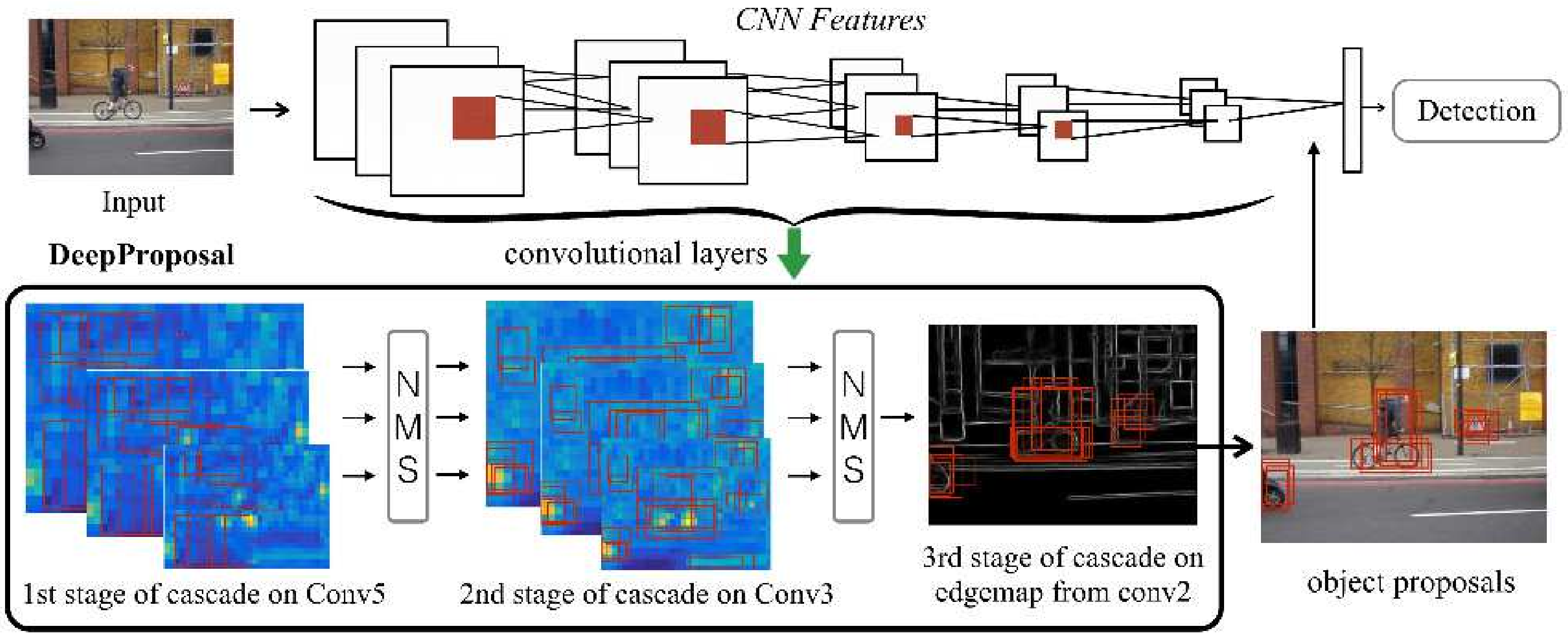


[Cimpoi et al., CVPR'15] [Ghodrati et al., ICCV'15]

Background (con't)

Utilities of pre-trained deep models

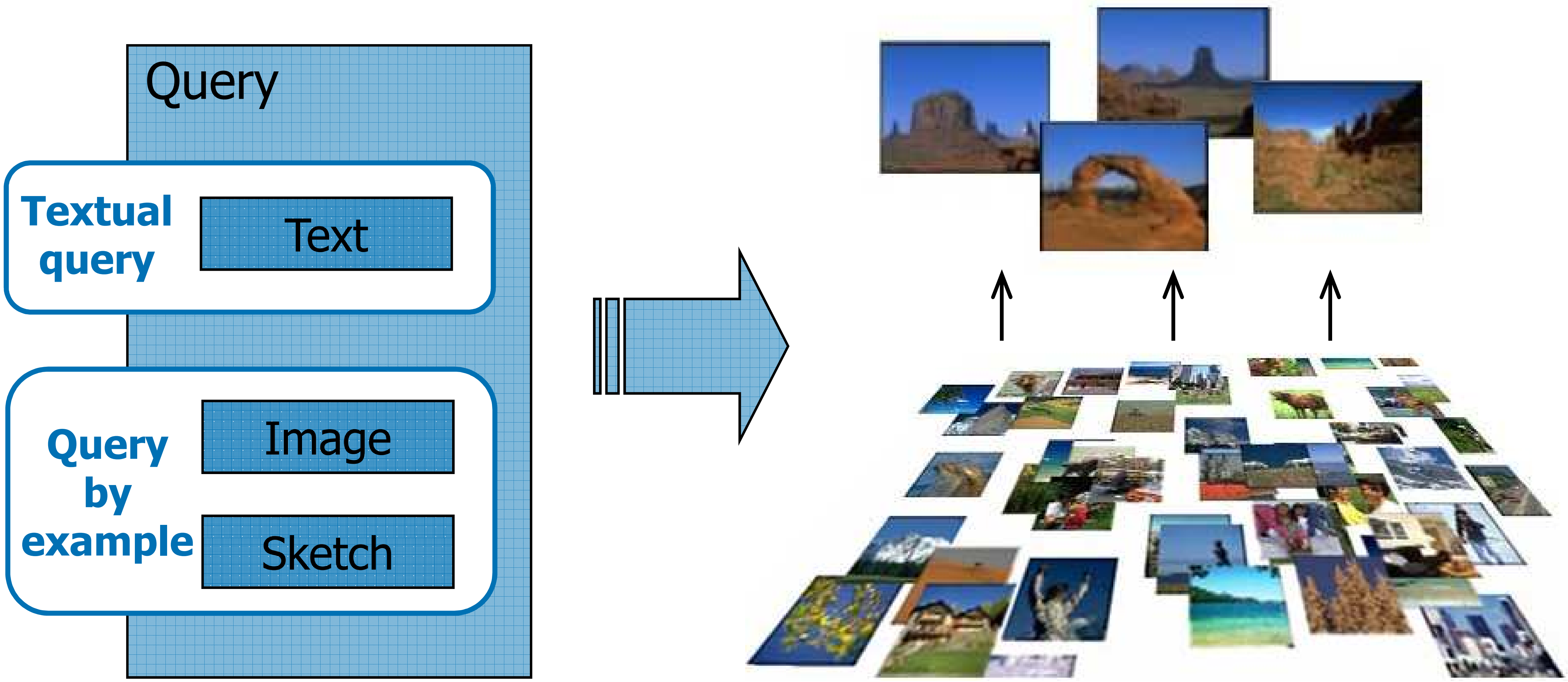
Still require further annotations in the new domain (e.g., image labels)



[Cimpoi et al., CVPR'15] [Ghodrati et al., ICCV'15]

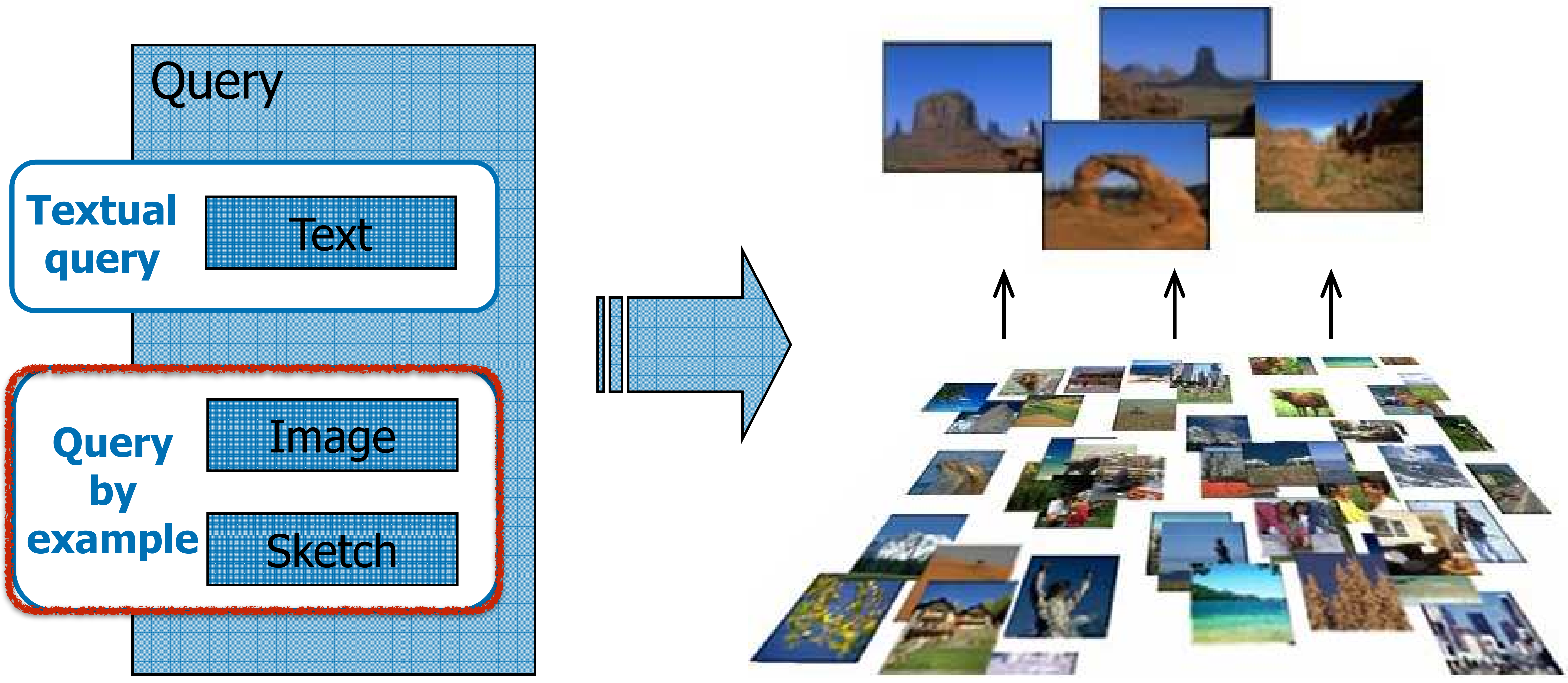
Background (con't)

Image Retrieval (IR)



Background (con't)

Image Retrieval (IR)



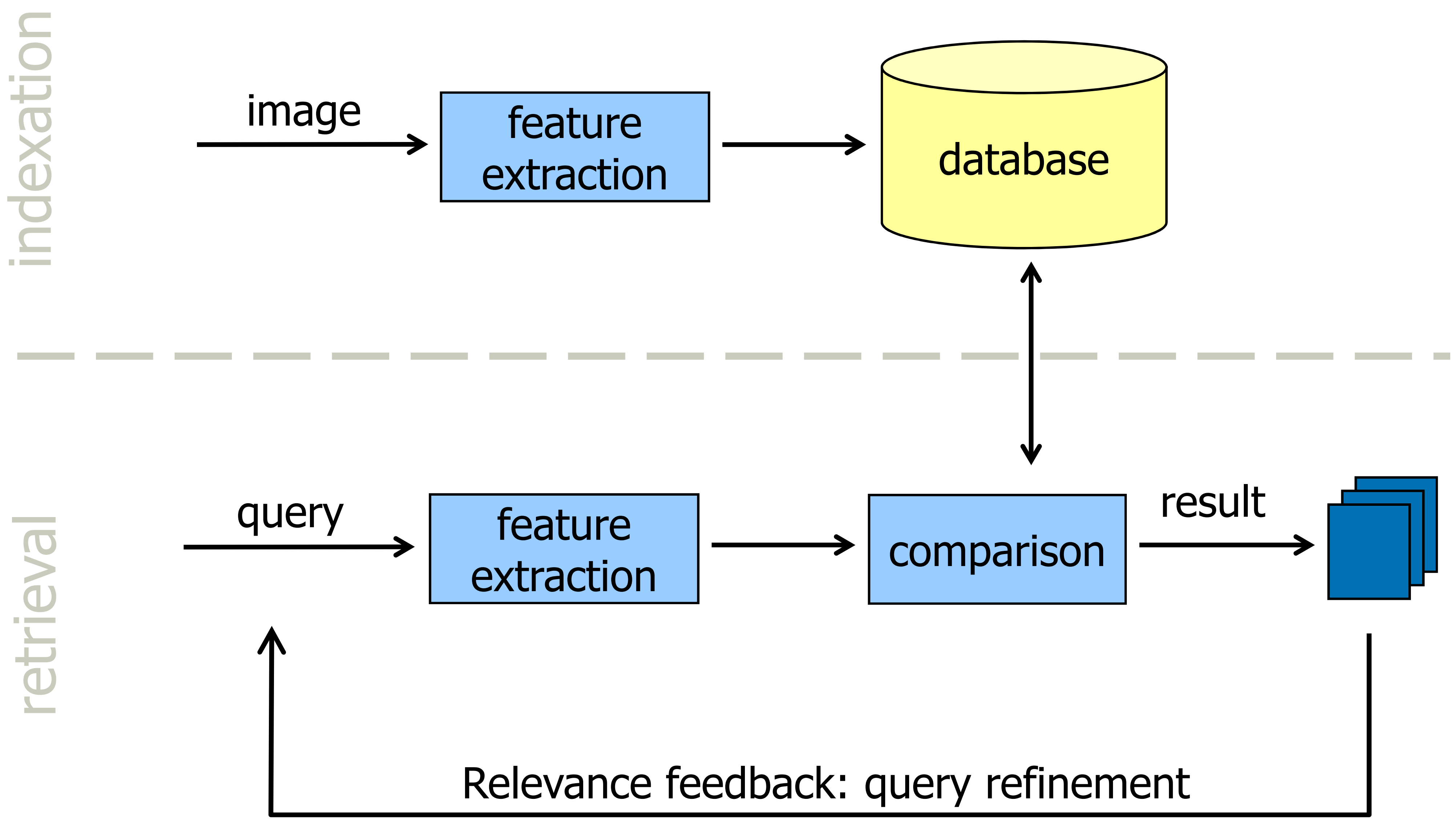
Background (con't)

Content-Based Image Retrieval (CBIR)

- Huge amounts of images are everywhere: how to manage this data?
- "A picture is worth thousand words."
- Automatic generation of textual annotations for a wide spectrum of images is not feasible.
- Annotating images manually is a cumbersome and expensive task for large image databases.
- Manual annotations are often subjective, context-sensitive and incomplete.
- …

Background (con't)

Common components of CBIR systems

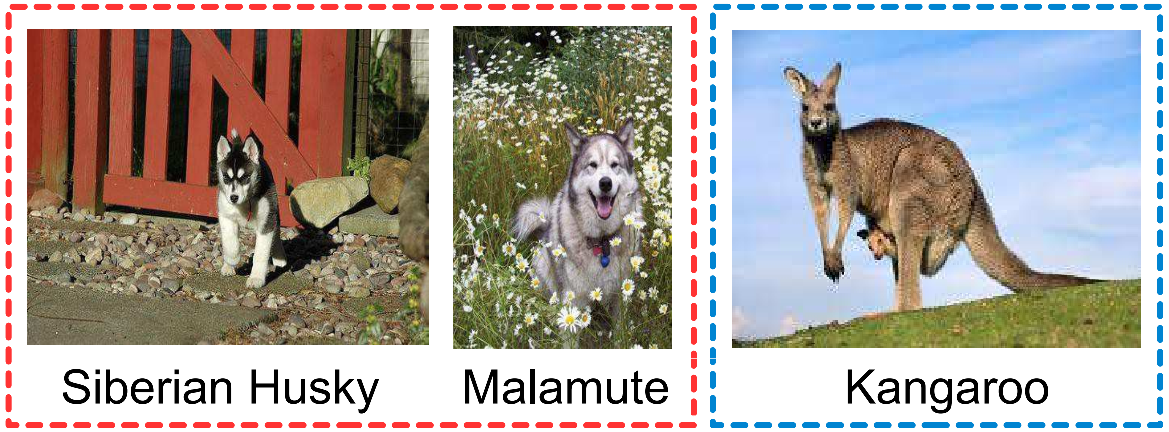


Related work

Deep learning for image retrieval

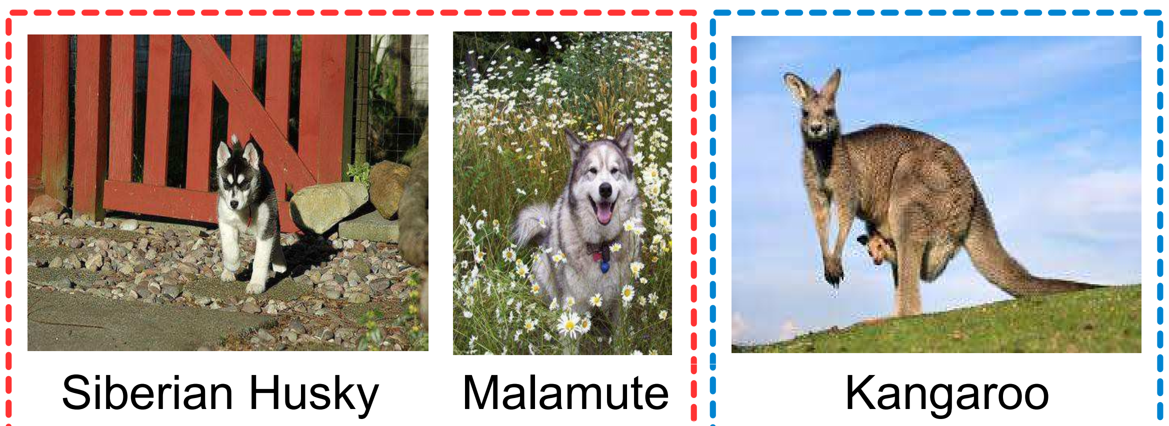
Related work (con't)

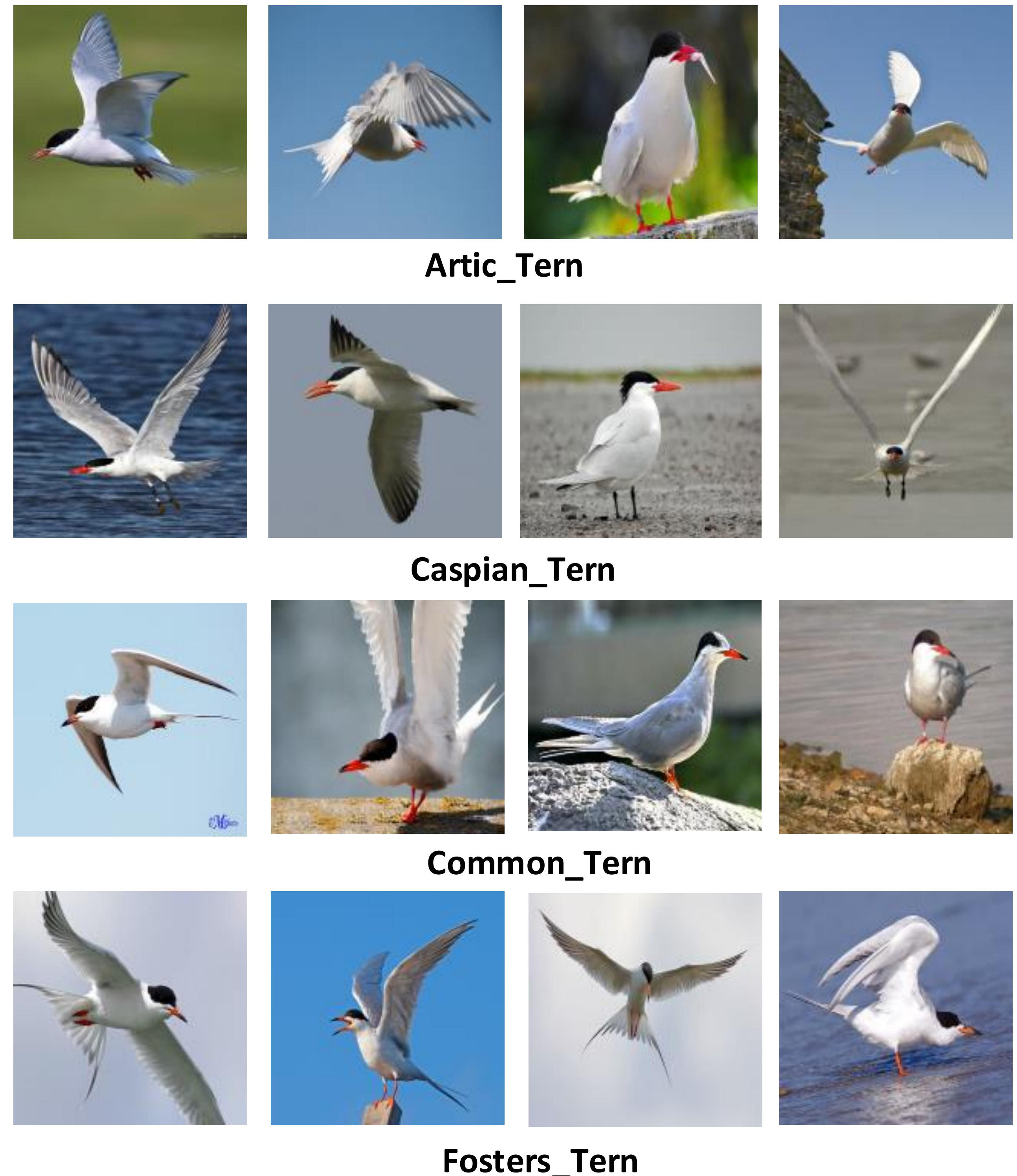
Fine-grained image tasks



Related work (con't)

Fine-grained image tasks

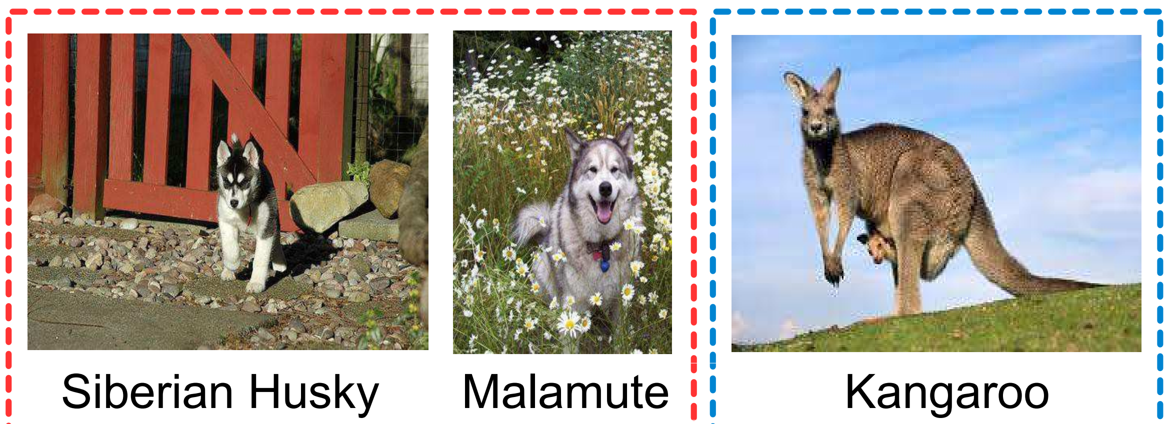


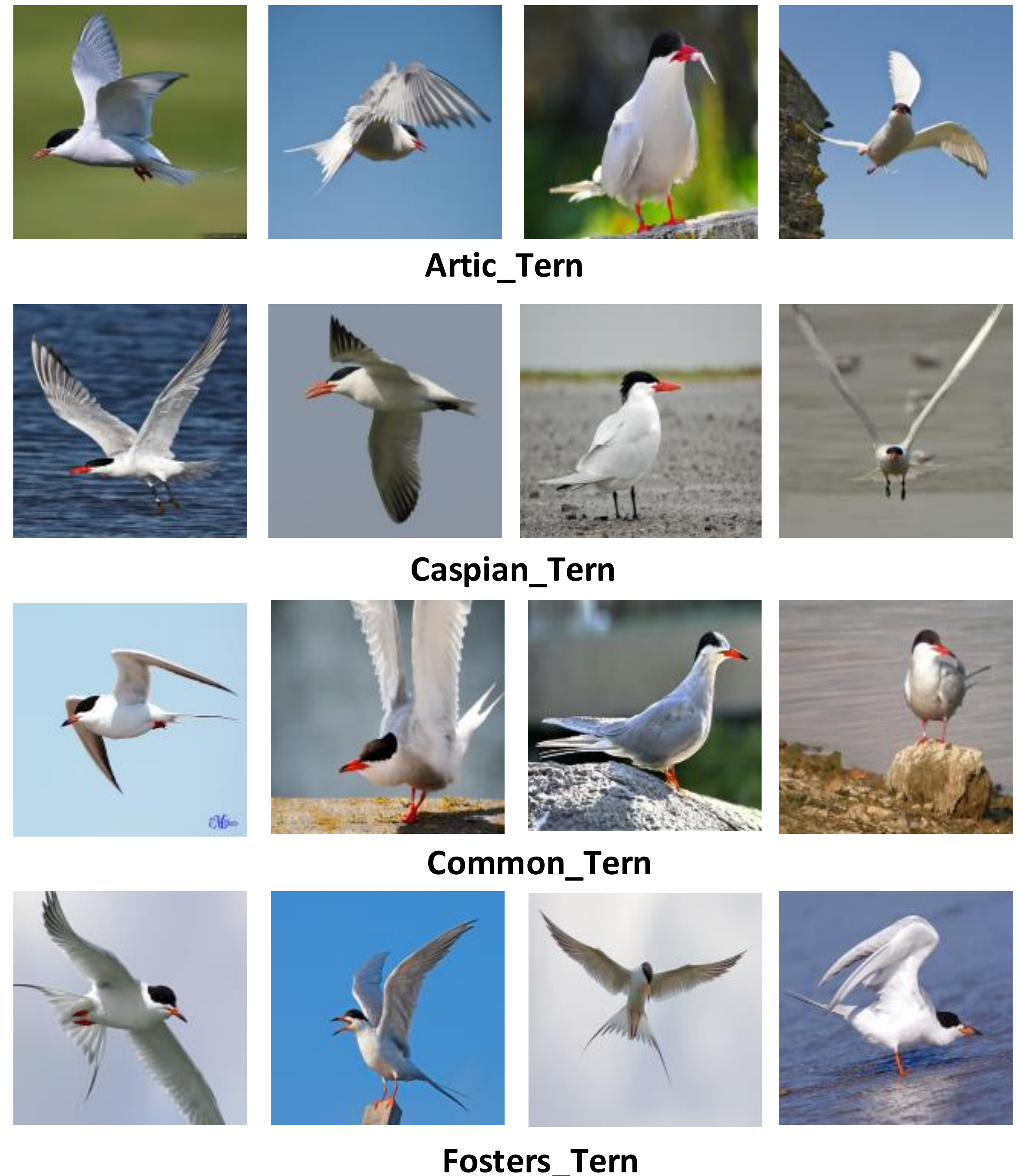


Related work (con't)

LAMDA
Learning And Mining from DatA

Fine-grained image tasks

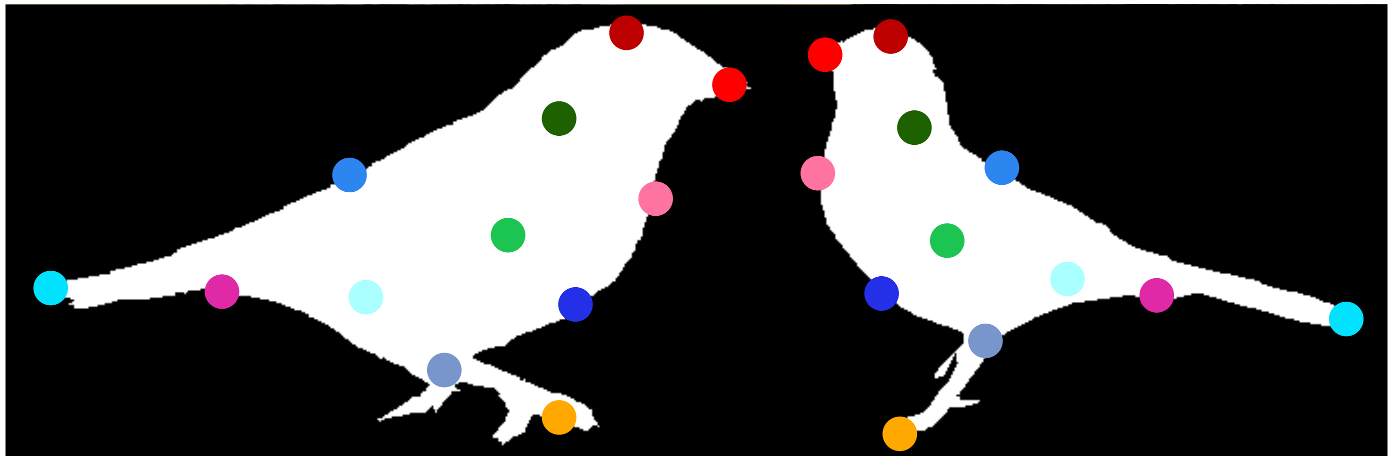




Related work (con't)

Fine-grained classification (supervised or weakly supervised)

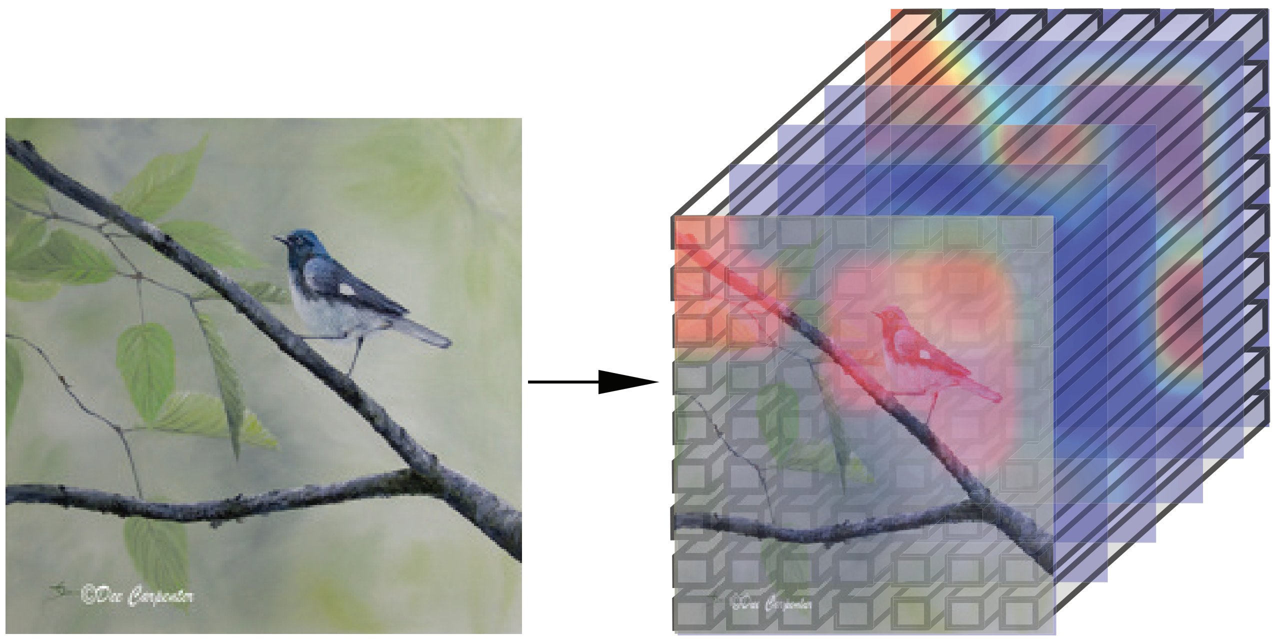
Bounding boxes



Part annotations

The proposed method

Notations



(a) Input image

(b) Convolutional activation tensor

$h \times w \times d$

Feature maps:

2-D feature maps $S = \{S_n\}$ $(n = 1, \ldots, d)$

Descriptors:

$X = \{\boldsymbol{x}_{(i,j)}\}$

The proposed method (con't)

Selective Convolutional Descriptor Aggregation (SCDA)

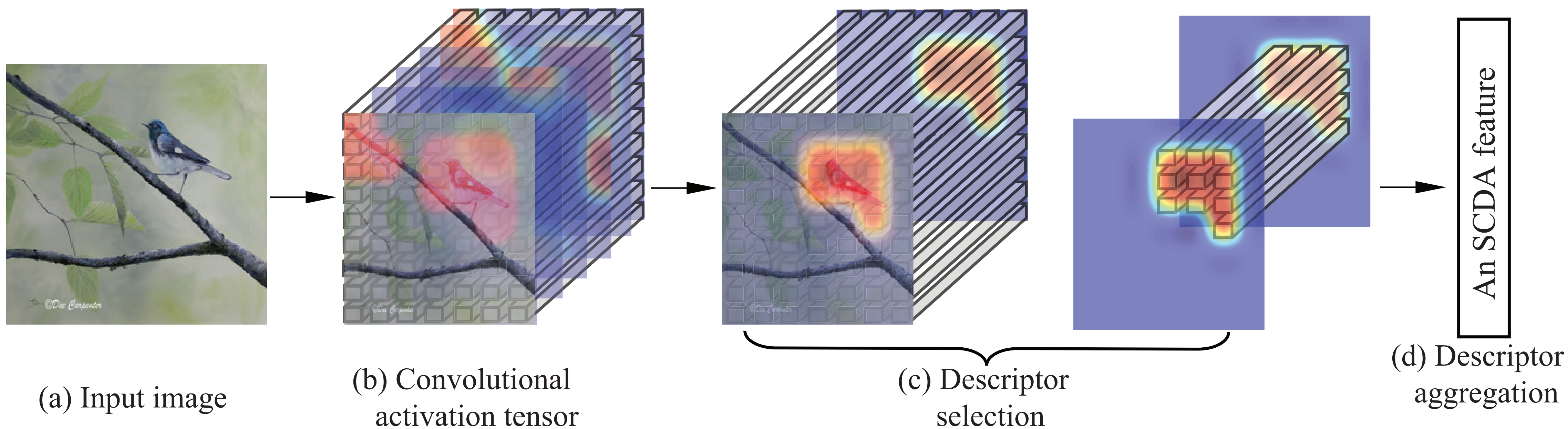
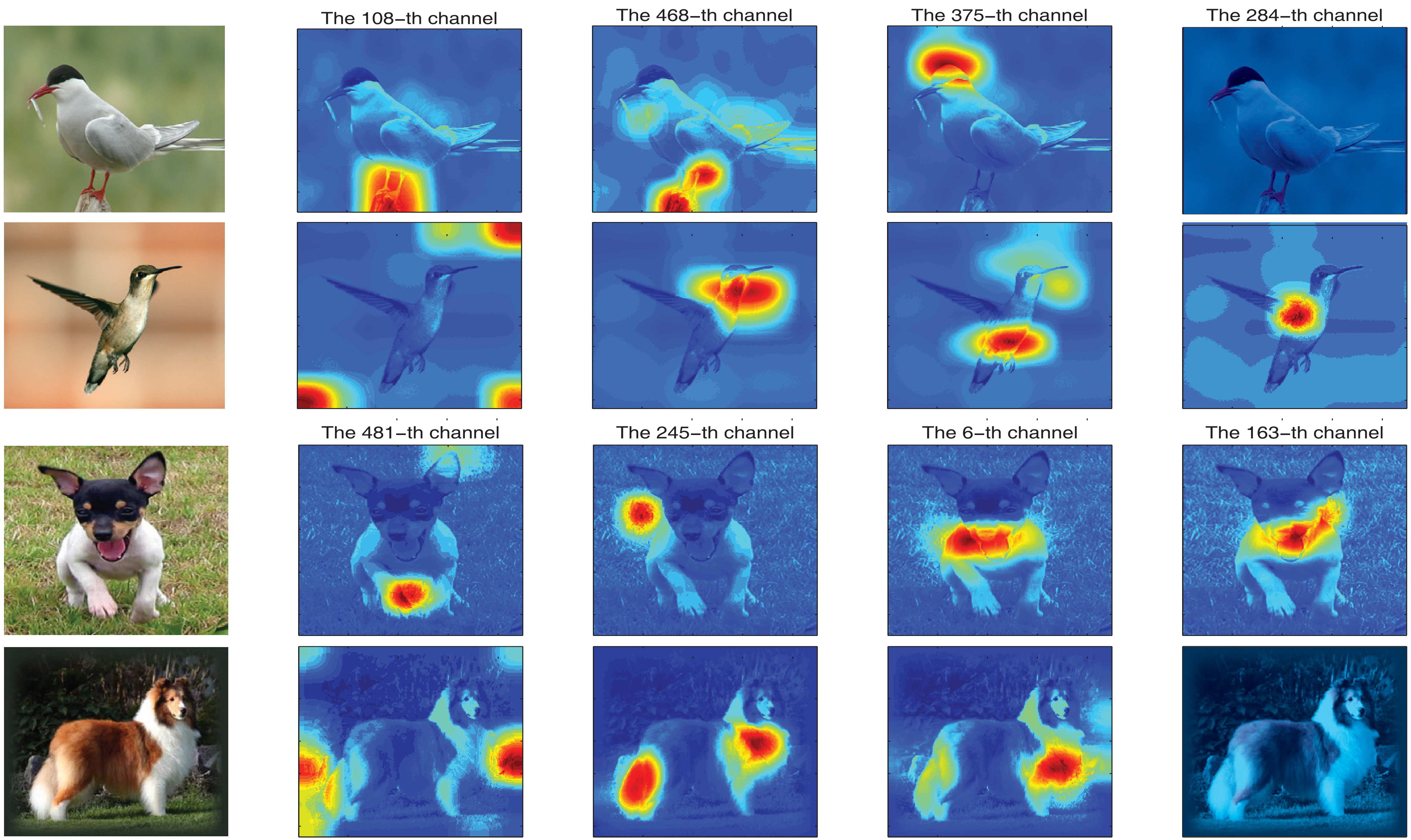


Figure 1. Pipeline of the proposed SCDA method. (Best viewed in color.)

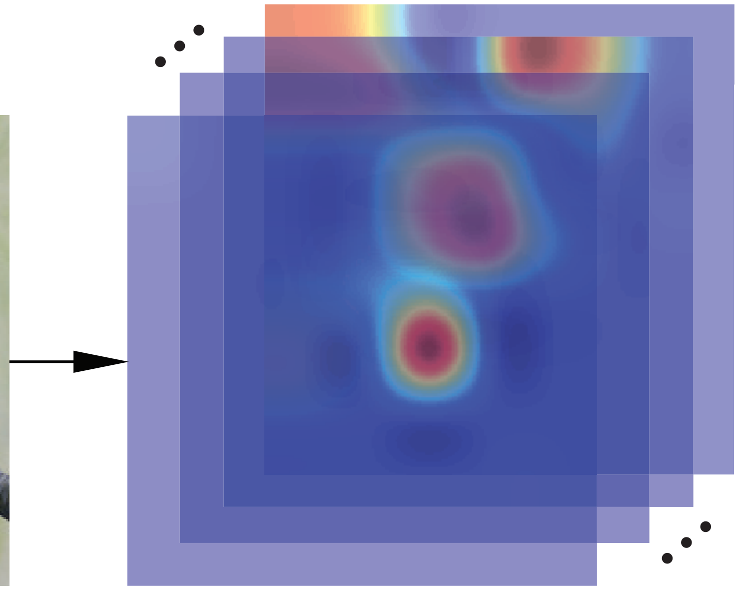
The proposed method (con't)



The proposed method (con't)

Obtaining the activation map by summarizing feature maps

(a) Input image

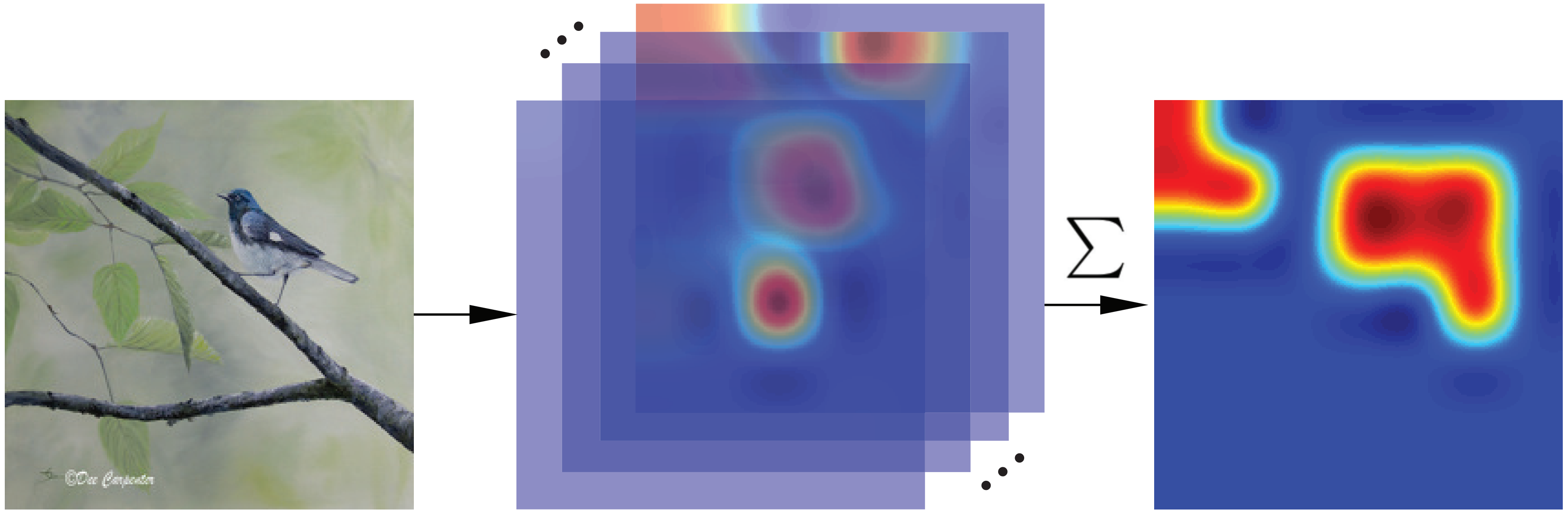


(b) Feature maps

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

The proposed method (con't)

Obtaining the activation map by summarizing feature maps



(a) Input image (b) Feature maps (c) Activation map

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

The proposed method (con't)

Visualization of the mask map M

(a) Visualization of the mask map M

The proposed method (con't)

Visualization of the mask map M

(a) Visualization of the mask map M

(b) Visualization of the mask map $\widetilde{M}$

The proposed method (con't)

Selecting useful deep convolutional descriptors

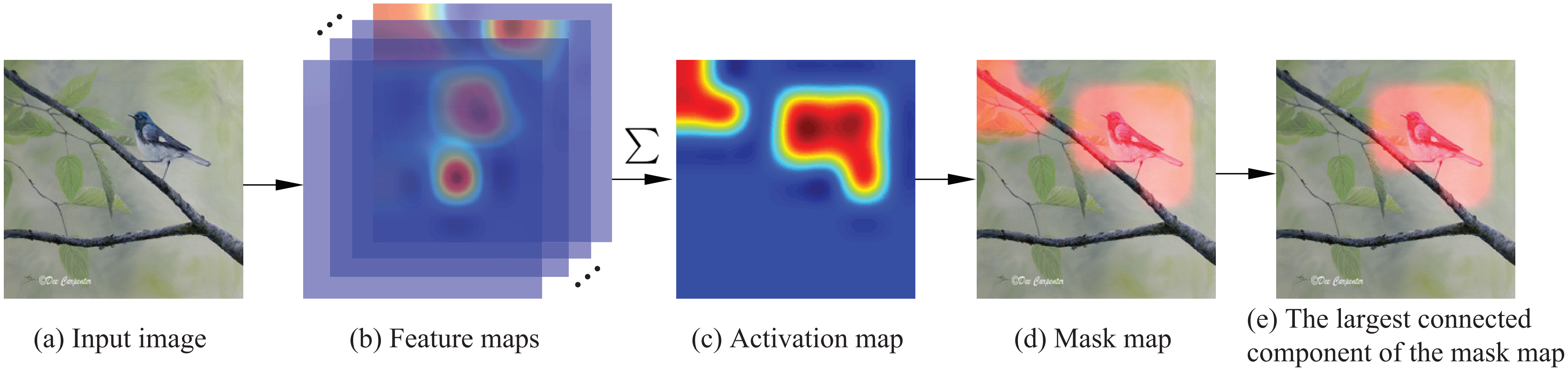
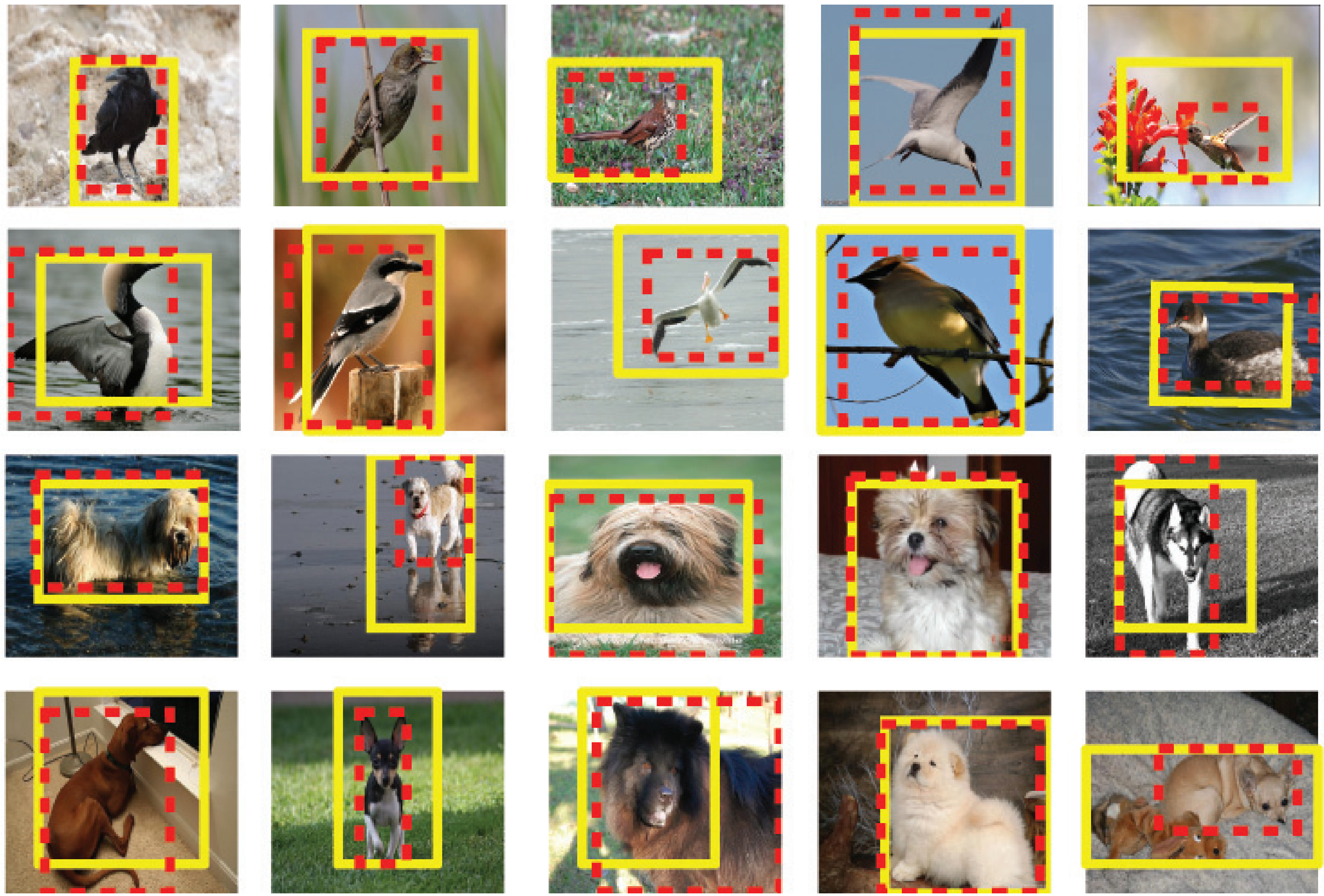


Figure 4. Selecting useful deep convolutional descriptors. (Best viewed in color.)

The proposed method (con't)

Learning And Mining from DatA

Qualitative evaluation



The proposed method (con't)

Quantitative evaluation

Table 1. Comparison of object localization performance on two fine-grained datasets.

Dataset	Method	Train phase		Test phase		Head	Torso	Whole-object
		BBox	Parts	BBox	Parts			
CUB200-2011	Strong DPM [29]	✓	✓	✓		43.49%	75.15%	N/A
	Part-based R-CNN with BBox [4]	✓	✓	✓		68.19%	79.82%	N/A
	Deep LAC [5]	✓	✓	✓		74.00%	96.00%	N/A
	Part-based R-CNN [4]	✓	✓			61.42%	70.68%	N/A
	Ours					N/A	N/A	**76.79%**
Stanford Dogs	**Ours**					N/A	N/A	**78.86%**

The proposed method (con't)

Aggregating convolutional descriptors

- **VLAD** [14] uses k-means to find a codebook of K centroids $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$ and maps $\boldsymbol{x}_{(i,j)}$ into a single vector $\boldsymbol{v}_{(i,j)} = \begin{bmatrix}\mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{x}_{(i,j)} - \boldsymbol{c}_k & \ldots & \mathbf{0}\end{bmatrix} \in \mathcal{R}^{K\times d}$, where $\boldsymbol{c}_k$ is the closest centroid to $\boldsymbol{x}_{(i,j)}$. The final representation is $\sum_{i,j} \boldsymbol{v}_{(i,j)}$.

- **Fisher Vector** [15]: FV is similar to VLAD, but uses a soft assignment (i.e., Gaussian Mixture Model) instead of using k-means. Moreover, FV also includes second-order statistics.[2]

- **Pooling approaches**. We also try two traditional pooling approaches, i.e., max-pooling and average-pooling, to aggregate the deep descriptors.

The proposed method (con't)

Comparing difference encoding or pooling methods

Approach	Dimension	*CUB200-2011*		*Stanford Dogs*	
		top1	top5	top1	top5
VLAD	1,024	55.92%	62.51%	69.28%	74.43%
Fisher Vector	2,048	52.04%	59.19%	68.37%	73.74%
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%**

The proposed method (con't)

Comparing difference encoding or pooling methods

Approach	Dimension	*CUB200-2011*		*Stanford Dogs*	
		top1	top5	top1	top5
VLAD	1,024	55.92%	62.51%	69.28%	74.43%
Fisher Vector	2,048	52.04%	59.19%	68.37%	73.74%
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%**

SCDA

The proposed method (con't)

Learning And Mining from DatA

Multiple layer ensemble

(a) M of Pool5 (b) $\widetilde{M}$ of Pool5 (c) M of Relu5_2 (d) $\widetilde{M}$ of Relu5_2

Figure 6. The mask map and its corresponding largest connected component of different CNN layers. (The figure is best viewed in color.)

The proposed method (con't)

Multiple layer ensemble

(a) M of Pool5 (b) $\widetilde{M}$ of Pool5 (c) M of Relu5_2 (d) $\widetilde{M}$ of Relu5_2

Figure 6. The mask map and its corresponding largest connected component of different CNN layers. (The figure is best viewed in color.)

$$\mathrm{SCDA}^{+} \leftarrow \left[\mathrm{SCDA}_{\mathrm{pool}_5},\ \alpha \times \mathrm{SCDA}_{\mathrm{relu}_{5\_2}}\right]$$

The proposed method (con't)

Multiple layer ensemble

(a) M of Pool5 (b) $\widetilde{M}$ of Pool5 (c) M of Relu5_2 (d) $\widetilde{M}$ of Relu5_2

Figure 6. The mask map and its corresponding largest connected component of different CNN layers. (The figure is best viewed in color.)

$$\mathrm{SCDA}^{+} \leftarrow \left[\mathrm{SCDA}_{\mathrm{pool}_5},\ \alpha \times \mathrm{SCDA}_{\mathrm{relu}_{5\_2}}\right]$$

$$\mathrm{SCDA\_flip}^{+}$$

The proposed method (con't)

Learning And Mining from DatA

Key advantages and main contributions:

- ☑ We propose a simple yet effective approach to **localize** the main object. This localization is **unsupervised**, **without utilizing bounding boxes, image labels, object proposals, or additional learning**. SCDA selects only useful deep descriptors and removes background or noise, which benefits the retrieval task.

- ☑ As shown in experiments, the compressed SCDA feature has **stronger correspondence to visual attributes** (even subtle ones) than the deep activations, which might explain the success of SCDA for fine-grained tasks.

Experiments

Datasets

- ***CUB200-2011***: 200 birds classes, 11,788 images;
- ***Stanford Dogs***: 120 dogs classes, 20,580 image;
- ***Oxford Flowers:*** 102 flowers classes, 8,189 images;
- ***Oxford-IIIT Pets***: 37 dogs or cats classes, 7,349 images.

Experiments (con't)

LAMDA
Learning And Mining from DatA

Retrieval performance:

Method	Dimension	*CUB200-2011*		*Stanford Dogs*		*Oxford Flowers*		*Oxford Pets*	
		top1	top5	top1	top5	top1	top5	top1	top5
fc_8_im	4,096	39.90%	48.10%	66.51%	72.69%	55.37%	60.37%	82.26%	86.02%
fc_8_gtBBox	4,096	47.55%	55.34%	70.41%	76.61%	–	–	–	–
fc_8_predBBox	4,096	45.24%	53.05%	68.78%	74.09%	57.16%	62.24%	85.55%	88.47%
$Pool_5$	1,024	57.54%	63.66%	69.98%	75.55%	70.73%	74.05%	85.09%	87.74%
selectFV	2,048	52.04%	59.19%	68.37%	73.74%	70.47%	73.60%	85.04%	87.09%
selectVLAD	1,024	55.92%	62.51%	69.28%	74.43%	73.62%	76.86%	85.50%	87.94%
SPoC (w/o cen.)	256	34.79%	42.54%	48.80%	55.95%	71.36%	74.55%	60.86%	67.78%
SPoC (with cen.)	256	39.61%	47.30%	48.39%	55.69%	65.86%	70.05%	64.05%	71.22%
CroW	256	53.45%	59.69%	62.18%	68.33%	73.67%	76.16%	76.34%	80.10%
SCDA	1,024	59.72%	65.79%	74.86%	79.24%	75.13%	77.70%	87.63%	89.26%
$SCDA^+$	2,048	59.68%	65.83%	74.15%	78.54%	75.98%	78.49%	87.99%	89.49%
$SCDA\_flip^+$	4,096	**60.65%**	**66.75%**	**74.95%**	**79.27%**	**77.56%**	**79.77%**	**88.19%**	**89.65%**

Experiments (con't)

Retrieval performance:

Method	Dimension	*CUB200-2011*		*Stanford Dogs*		*Oxford Flowers*		*Oxford Pets*	
		top1	top5	top1	top5	top1	top5	top1	top5
fc_8_im	4,096	39.90%	48.10%	66.51%	72.69%	55.37%	60.37%	82.26%	86.02%
fc_8_gtBBox	4,096	47.55%	55.34%	70.41%	76.61%	–	–	–	–
fc_8_predBBox	4,096	45.24%	53.05%	68.78%	74.09%	57.16%	62.24%	85.55%	88.47%
$Pool_5$	1,024	57.54%	63.66%	69.98%	75.55%	70.73%	74.05%	85.09%	87.74%
selectFV	2,048	52.04%	59.19%	68.37%	73.74%	70.47%	73.60%	85.04%	87.09%
selectVLAD	1,024	55.92%	62.51%	69.28%	74.43%	73.62%	76.86%	85.50%	87.94%
SPoC (w/o cen.)	256	34.79%	42.54%	48.80%	55.95%	71.36%	74.55%	60.86%	67.78%
SPoC (with cen.)	256	39.61%	47.30%	48.39%	55.69%	65.86%	70.05%	64.05%	71.22%
CroW	256	53.45%	59.69%	62.18%	68.33%	73.67%	76.16%	76.34%	80.10%
SCDA	1,024	59.72%	65.79%	74.86%	79.24%	75.13%	77.70%	87.63%	89.26%
$SCDA^+$	2,048	59.68%	65.83%	74.15%	78.54%	75.98%	78.49%	87.99%	89.49%
$SCDA\_flip^+$	4,096	**60.65%**	**66.75%**	**74.95%**	**79.27%**	**77.56%**	**79.77%**	**88.19%**	**89.65%**

Experiments (con't)

LAMDA Learning And Mining from DatA

Post-processing

SCDA_flip$^+$	4,096	**60.65%**	**66.75%**	**74.95%**	**79.27%**	**77.56%**	**79.77%**	**88.19%**	**89.65%**

Table 4. Comparison of different compression methods on "SCDA_flip$^+$".

Method	Dimension	*CUB200-2011*		*Stanford Dogs*		*Oxford Flowers*		*Oxford Pets*	
		top1	top5	top1	top5	top1	top5	top1	top5
PCA	256	60.48%	66.55%	74.63%	79.09%	76.38%	79.32%	87.82%	89.75%
	512	60.37%	66.78%	74.76%	79.27%	77.15%	79.50%	87.46%	89.71%
SVD	256	60.34%	66.57%	**74.79%**	**79.27%**	76.79%	79.32%	**87.84%**	**89.79%**
	512	60.41%	66.82%	74.72%	79.26%	77.10%	79.48%	87.41%	89.72%
SVD+whitening	256	**62.29%**	**68.16%**	71.57%	76.68%	80.74%	82.42%	85.47%	87.99%
	512	62.13%	68.13%	71.07%	76.06%	**81.44%**	**82.82%**	85.23%	87.62%

Experiments (con't)

Learning And Mining from DatA

Experiments (con't)

Quality demonstration of the SCDA feature

Experiments (con't)

Classification accuracy

Table 5. Comparison of classification accuracy on four fine-grained datasets. The "details" column is a short description of the implementation details. ("f.t." stands for "fine-tune", and "h.flip" is short for "horizontal flip".)

Method	Train phase		Test phase		Details	Dim.	*Birds*	*Dogs*	*Flowers*	*Pets*
	BBox	Parts	BBox	Parts						
PB R-CNN with BBox [4]	✓	✓	✓		Alex-Net; f.t. on whole images and parts; with crops	12,288	76.4%	–	–	–
Deep LAC [5]	✓	✓	✓		Alex-Net; f.t. on whole images and parts; with crops	12,288	80.3%	–	–	–
PB R-CNN [4]	✓	✓			Alex-Net; f.t. on whole images and parts; with crops	12,288	73.9%	–	–	–
Two-Level [6]					VGG-16; f.t. with part proposals	16,384	77.9%	–	–	–
Weakly supervised FG [9]					VGG-16; f.t. with h.flip	262,144	79.3%	80.4%	–	–
Constellations [7]					VGG-19; f.t. with h.flip; with part proposals	208,896	81.0%	68.6%[1]	95.3%	91.6%
Bilinear [8]					VGG-19 and VGG-M; training with h.flip	262,144	84.0%	–	–	–
Spatial Transformer Net [34]					Inception architecture; training with h.flip and crops	4,096	84.1%	–	–	–
Ours					**VGG-16; f.t. with h.flip; w/o crops**	**4,096**	**80.5%**	**78.7%**	**92.1%**	**91.0%**

Conclusions and future work

Conclusions

- solely using a CNN model pre-trained on non-fine-grained tasks
- the proposed SCDA: unsupervised and without additional learning
- satisfactory retrieval results and corresponding to semantic visual attributes

Conclusions and future work

Conclusions

- solely using a CNN model pre-trained on non-fine-grained tasks
- the proposed SCDA: unsupervised and without additional learning
- satisfactory retrieval results and corresponding to semantic visual attributes

Future work

- We consider including the selected descriptors' weights to find parts.
- We also want to explore the possibility of pre-trained models for more complicated vision tasks, e.g., object segmentation unsupervised.

Learning And Mining from DatA

Thank you!