Unsupervised Learning of Discriminative Attributes and Visual Representations

Chen Huang[1,2], Chen Change Loy[1], Xiaoou Tang[1]

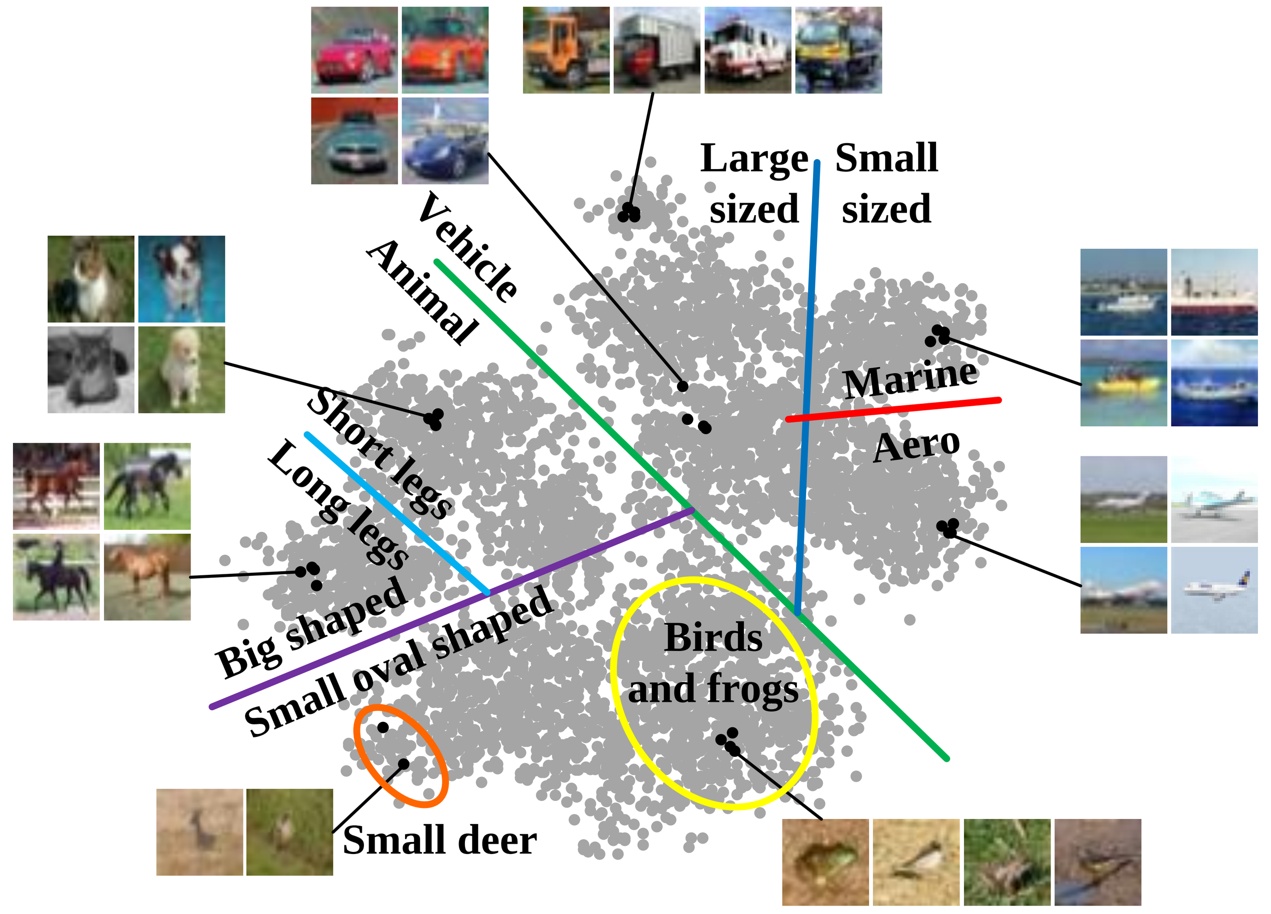
[1]The Chinese University of Hong Kong [2]SenseTime Group Limited

{chuang, ccloy, xtang}@ie.cuhk.edu.hk

CVPR 2016

1. Motivation

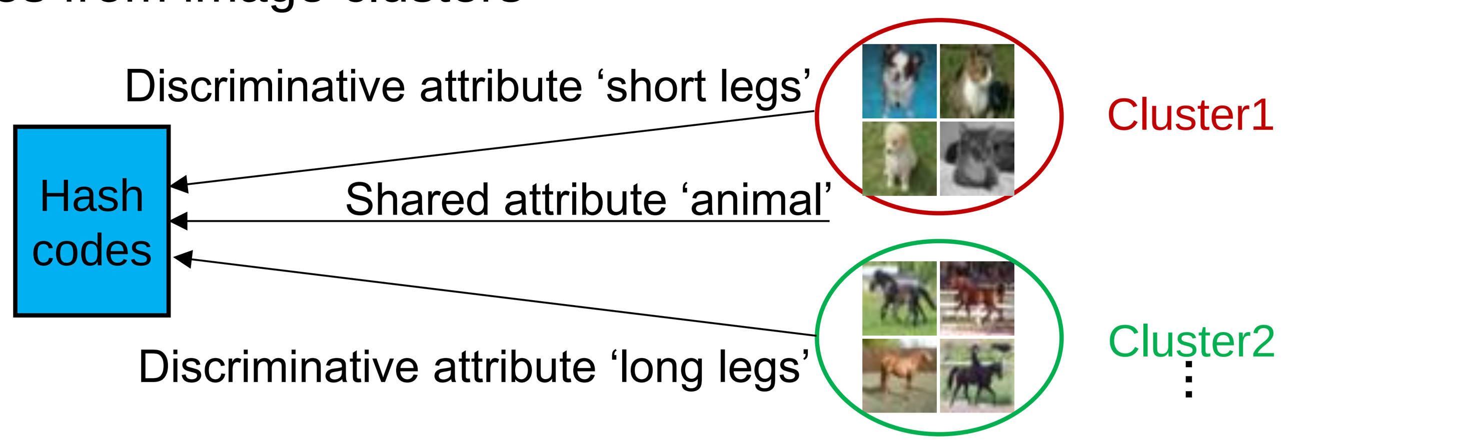
- Visual attributes offer useful mid-level cues
- Most **visual attribute** and **visual representation** learning methods are supervised by costly labels
- Goal: unsupervised learning of both directly from data



Unsupervisedly learned 2D feature space and attributes on CIFAR-10

2. Related Work & Main Idea

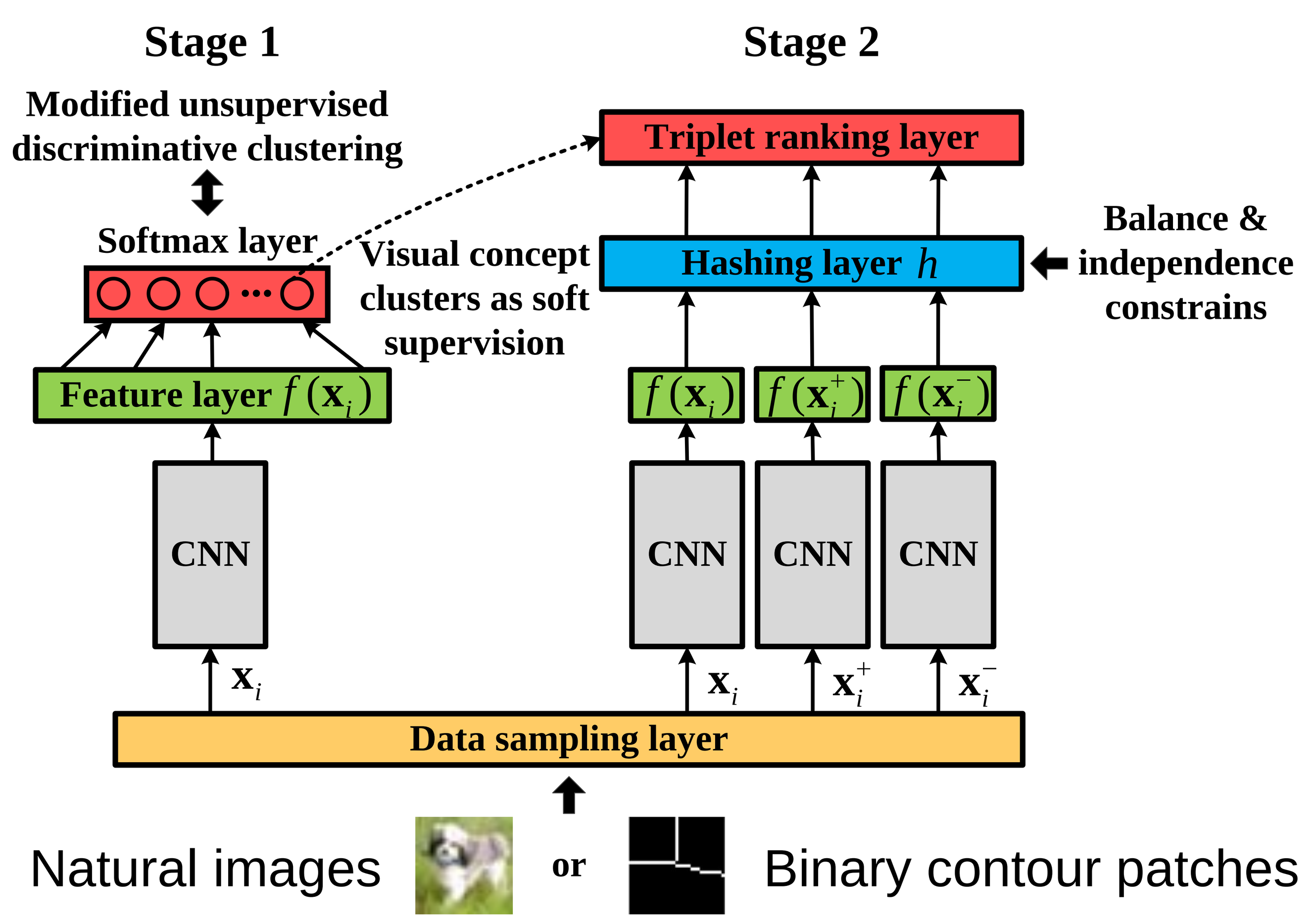
- Related work
  - "Unsupervised" **attribute** learning on the class basis Still supervised
  - Unsupervised **feature** learning by predicting within-image contexts, ranking patches from video tracks, etc. Attribute untouched
- Main idea
  - Learn to extract **shared** and **discriminative** binary hash codes as attributes from image clusters



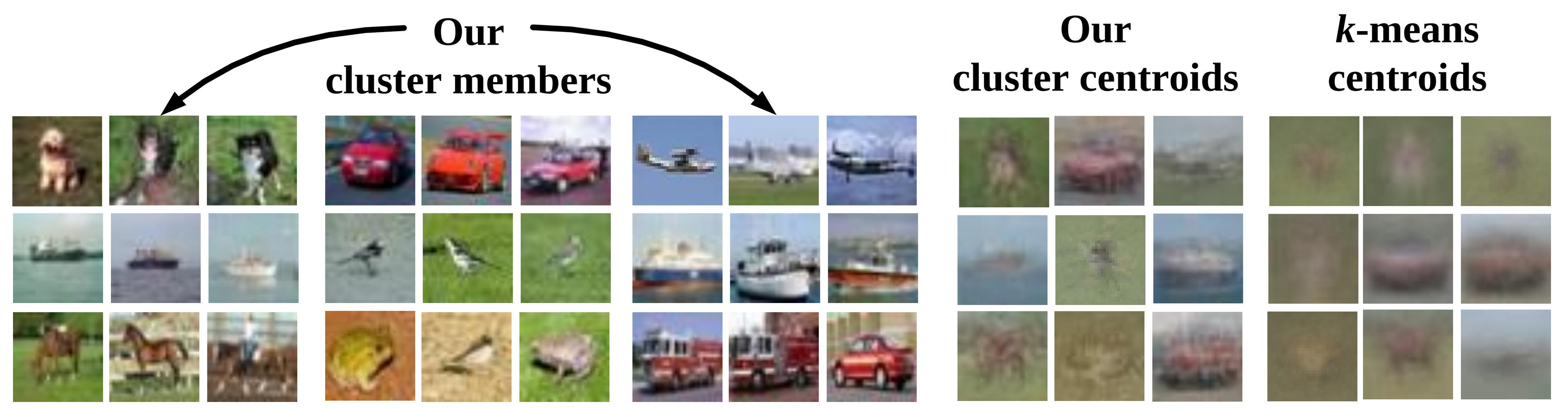
  - Note the learned attributes are not strictly "attributes" (more related to attribute hypothesis). But they still highly correlate with semantics.

3. Approach

- Two-stage pipeline



- Stage 1
  - Modify clustering algorithm [Singh, ECCV12]: cluster merging & augmentation
  - Alternating with CNN feature learning (softmax classification)



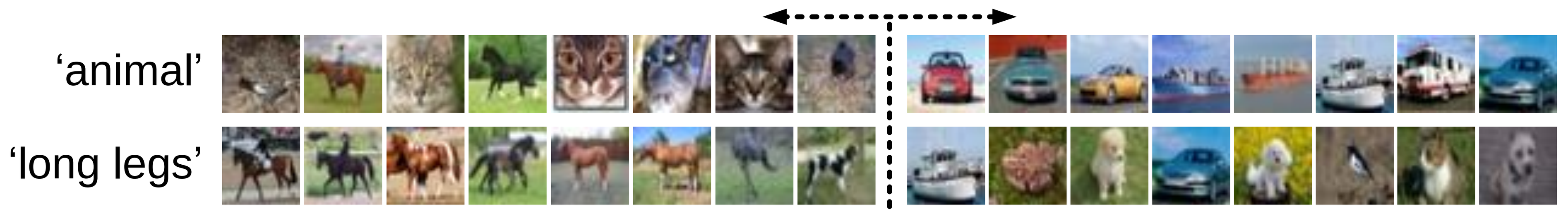
- Stage 2
  - Weakly-supervised hashing (K-bits): triplet ranking loss

$$\min \sum_i \varepsilon_i + \alpha tr\left[\mathbf{W}^T f(\mathbf{X}) f(\mathbf{X})^T \mathbf{W}\right] + \beta \left\|\mathbf{W}\mathbf{W}^T - \mathbf{I}\right\|_2^2 + \gamma \left\|\mathbf{W}\right\|_2^2,$$

$$s.t.: \quad \max\left(0, \rho + H\left(\mathbf{b}_i, \mathbf{b}_i^+\right) - H\left(\mathbf{b}_i, \mathbf{b}_i^-\right)\right) \le \varepsilon_i,$$

$$\forall i, \quad \mathbf{b}_i = h\left(x_i; \mathbf{W}\right) = 2\sigma\left(\mathbf{W}^T f(\mathbf{x}_i)\right) - 1, \quad \varepsilon_i \ge 0,$$

$$H\left(\mathbf{b}_i, \mathbf{b}_j\right) = \left(K - \mathbf{b}_i^T \mathbf{b}_j\right)/2, \quad \rho = K/2$$

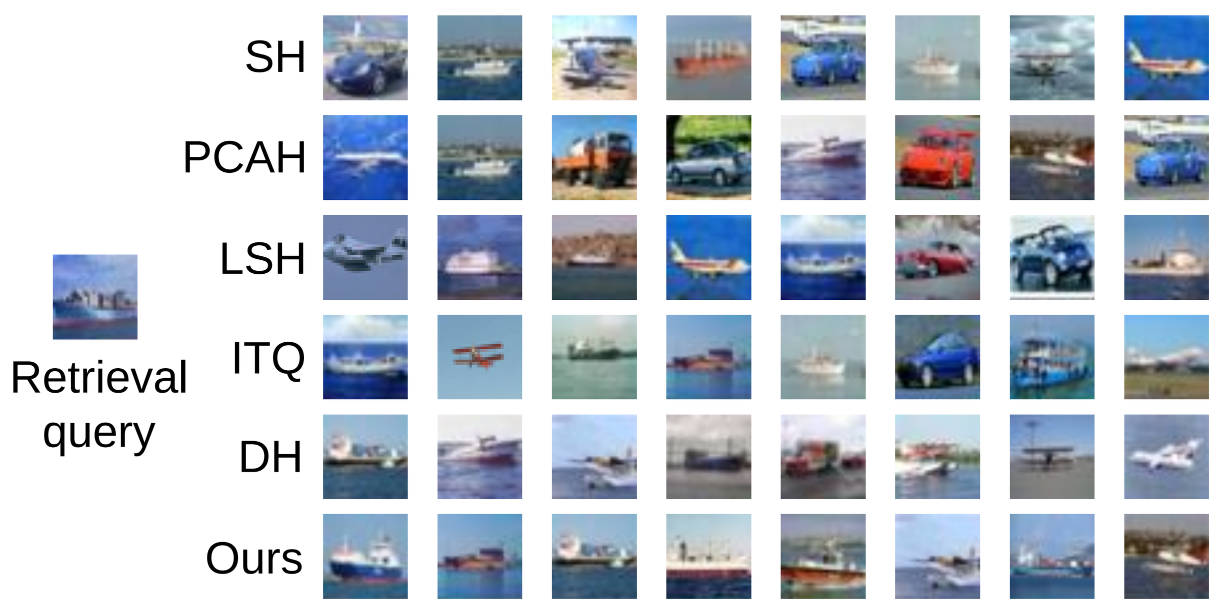


4. Results

- Unsupervised pre-training for Fast R-CNN **detection** (%) on PASCAL VOC 2007

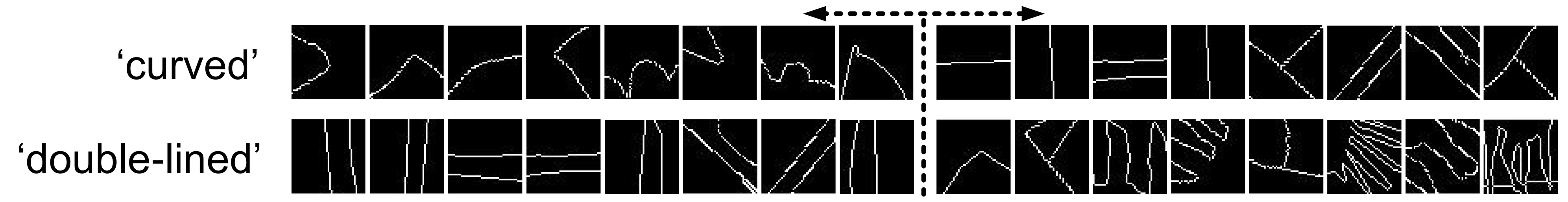
Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hors.	bike	Per.	plant	sh.p	sofa	train	tv	mAP
Doersch et al. +R-CNN	60.5	**66.5**	29.6	28.5	26.3	56.1	**70.4**	44.8	24.6	45.5	45.4	35.1	52.2	60.2	50	28.1	46.7	42.6	54.8	58.6	46.3
Doersch et al.	54.4	50.8	30.1	28.9	10.3	57.5	60.8	46.3	19.8	38.5	51.5	37.4	60.6	53	45.1	14.2	26	44.5	55.6	43.7	41.4
Wang and Gupta	53.9	53.9	30.5	29.6	10.8	56	59	46.1	19.6	45.7	43.9	41.6	65.6	58.6	48.2	17.4	34.8	41.2	64.5	46.5	43.3
Ours (K = 32 bits)	59.2	61.6	31.2	33.4	27	58.3	64.9	49.1	28.6	50.4	51.9	44.7	57.9	58.5	52.3	**29.6**	46.1	43.2	65.6	59.2	48.6
Ours (K = 64 bits)	**62.6**	63.7	**37.6**	34.9	**28.8**	57.6	66.2	51.6	**30.7**	**51.5**	48.5	47.1	55.6	59.8	49.2	27.5	**47.8**	41.4	63.6	**59.7**	49.3
CFN-9 (max)	61.5	64.3	36.4	**36.1**	20.8	**65.8**	69	**59.2**	30.3	50	**58.1**	**50.7**	**70.6**	**67.2**	**56**	22.7	44.7	**52.8**	**66.9**	52	**51.8**
ImageNet label	**65.1**	**70.3**	**53.6**	**41.6**	25.1	**69.3**	68.9	**68.8**	30.4	**63**	**62.3**	**63.3**	**72.7**	**64.5**	**57.1**	25.2	**50.6**	**54**	**70.1**	55.1	**56.5**

- Unsupervised learning for **image retrieval** (CIFAR-10) and **classification**

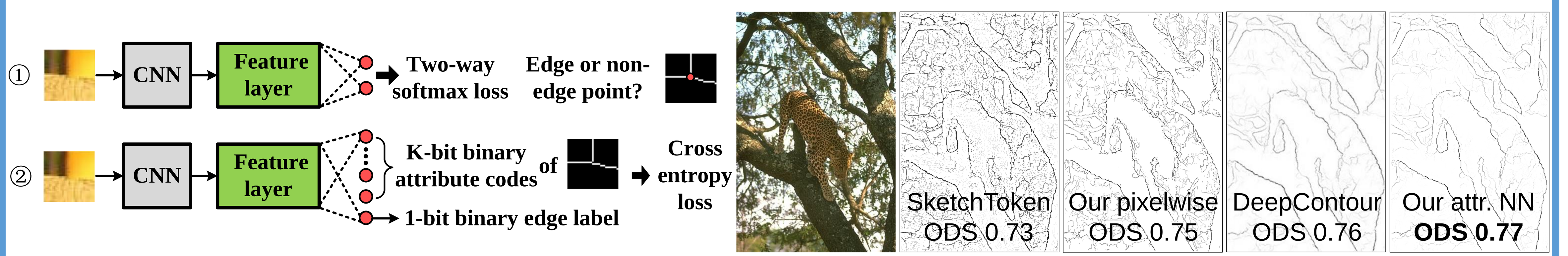


Method	STL-10	Caltech-101
Multi-way local pooling	-	77.3±0.6
Slowness on videos	61.0	74.6
HMP	64.5±1	-
Multipath HMP	-	82.5±0.5
View-Invariant k-means	63.7	-
Exemplar-CNN	75.4±0.3	87.2±0.6
Ours (K = 16 bits)	74.9±0.4	86.1±0.7
Ours (K = 32 bits)	76.3±0.4	87.8±0.5
Ours (K = 64 bits)	**76.8±0.3**	**89.4±0.5**
Supervised state of the art	70.1	**91.44**

- Unsupervised learning for **contour detection** on BSDS500
  - Unsupervised attribute learning from **binary contour patches**



  - Supervised learning from **natural image patches**



5. Conclusion

- **Unsupervised deep learning of visual attributes and representations** by unsupervised discriminative clustering and weakly-supervised hashing
- Capture strong semantic meanings, transferrable to other vision tasks