# Unsupervised Visual Representation Learning by Graph-based Consistent Constraints Supplementary Material

Dong Li[1], Wei-Chih Hung[2], Jia-Bin Huang[3],
Shengjin Wang[1⋆], Narendra Ahuja[3], and Ming-Hsuan Yang[2]

[1]Tsinghua University, [2]University of California, Merced,
[3]University of Illinois, Urbana-Champaign
https://sites.google.com/site/lidonggg930/feature-learning

## 1 Overview

In this supplementary material, we present four sets of additional results.

1. We show cycle detection results using the proposed unsupervised constraint mining approach on the large-scale ImageNet 2012 dataset as well as three image datasets with generic objects (CIFAR-10), fine-grained objects (CUB-200-2011), and scene classes (MIT indoor-67), respectively.
2. We show examples of easy negative image pairs (i.e., image pairs with large Euclidean distance in the feature space) and hard negative sample pairs (i.e., image pairs with large geodesic distance in the k-NN graph).
3. We show additional qualitative results and a detailed quantitative evaluation for the unsupervised feature learning task.
4. We report detailed quantitative results on three image classification datasets for the semi-supervised learning task.

## 2 Cycle detection

We show cycle detection results using the proposed unsupervised constraint mining method on the ImageNet (Fig. 2), CIFAR-10 (Fig. 3), CUB-200-2011 (Fig. 4), and MIT indoor-67 (Fig. 5) datasets. For all these results, we set the number of nearest neighbors $k = 4$ and the length of the cycle $n = 4$ in the mining algorithm. The results show that cycle consistency captures semantic information and helps mine positive image pairs with large appearance variations without using any manual annotations.

## 3 Negative mining

Fig. 6 shows several examples of negative mining results on the ImageNet dataset. Geodesic distance can discover hard negative image pairs with visually similar appearances. These image pairs provide valuable information for learning effective representations.

⋆ Corresponding author.

## 4 Evaluation on unsupervised feature learning

We demonstrate the quality of our learned representations on the ImageNet dataset with both qualitative and quantitative experiments. For qualitative evaluation, Fig. 1 shows the learned filters of the first convolutional layer. Fig. 7 and Fig. 8 show nearest neighbor retrieval results on the ImageNet 2012 *validation* set. With our method of unsupervised feature learning, we are able to obtain similar retrieval results as those obtained by the supervised pre-trained AlexNet for a variety of visual categories. This demonstrates that our learned representations capture semantic information although class labels are not available during the training stage.

For quantitative evaluation, we show in Table 1 the detailed classification performance on the PASCAL VOC 2007 *test* set using the learned representations on ImageNet. The results show that our unsupervised trained CNN model performs well for distinguishing visual categories on a particular dataset (51.8% mAP on VOC 2007) and has good generalization capability. With fine-tuning by the ground-truth image labels, the adapted representations achieve improved classification performance (56.5% mAP). We also present the precision-recall curves for each category in Fig. 9.

## 5 Evaluation on semi-supervised learning

We show detailed quantitative results on three image classification datasets in the semi-supervised setting. For the three datasets used, we randomly select $m$ images per class in the training set as the partial annotated data.

We show boosted classification performance by incorporating the mined constraints on CIFAR-10 in Table 2, CUB-200-2011 in Table 3, and MIT indoor-67 in Table 4. We achieve higher accuracy over the baseline which directly uses the limited labeled data to fine-tune the network. The results show that our method mines new effective constraints beyond annotations for learning better feature representations.

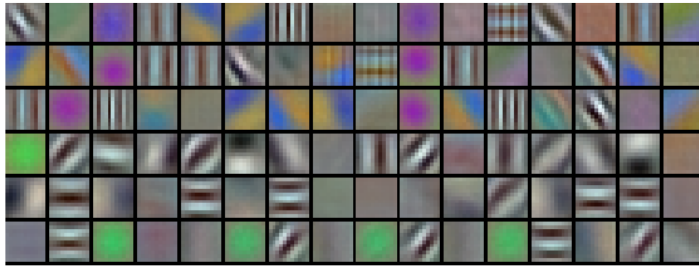


**Fig. 1.** Conv1 filters of the Siamese network learned in the unsupervised way.

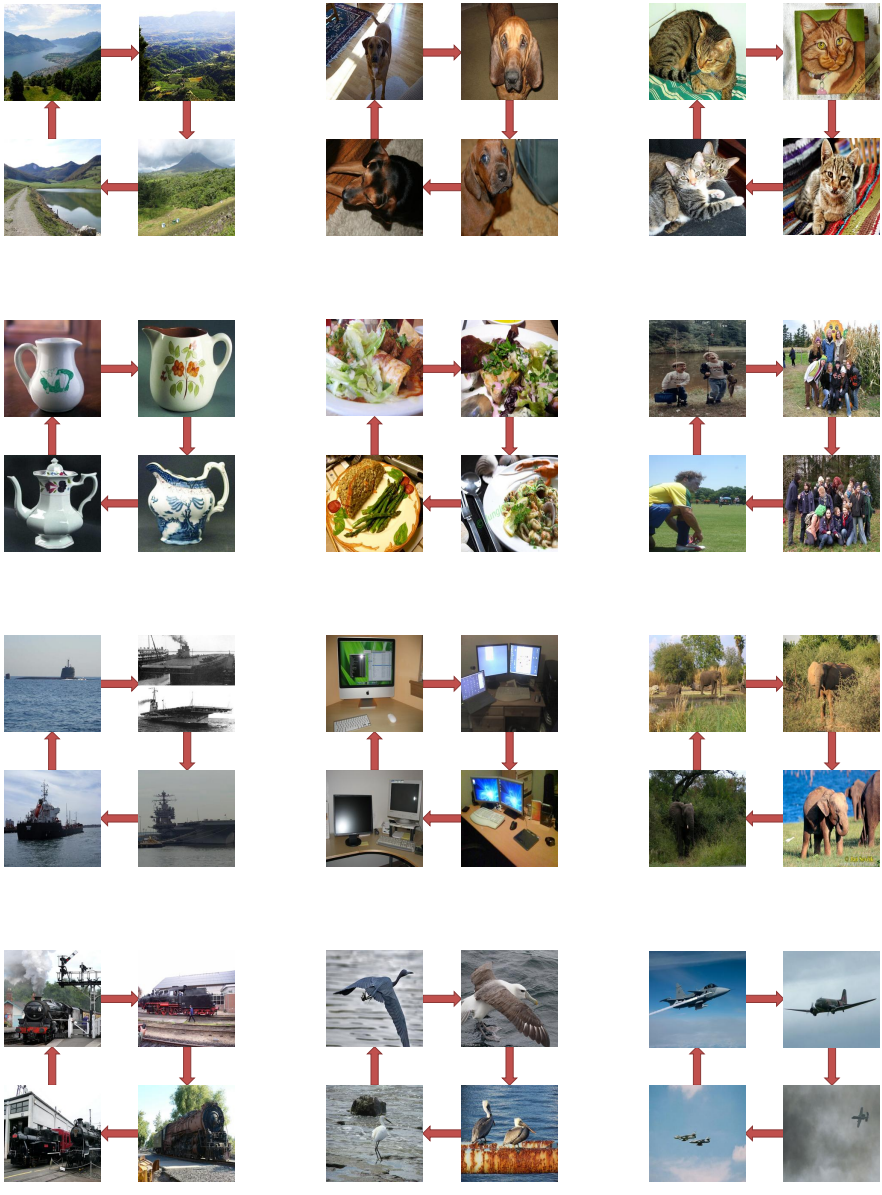


**Fig. 2.** Sample cycle detection results on the ImageNet dataset.

**Fig. 3.** Sample cycle detection results on the CIFAR-10 dataset.

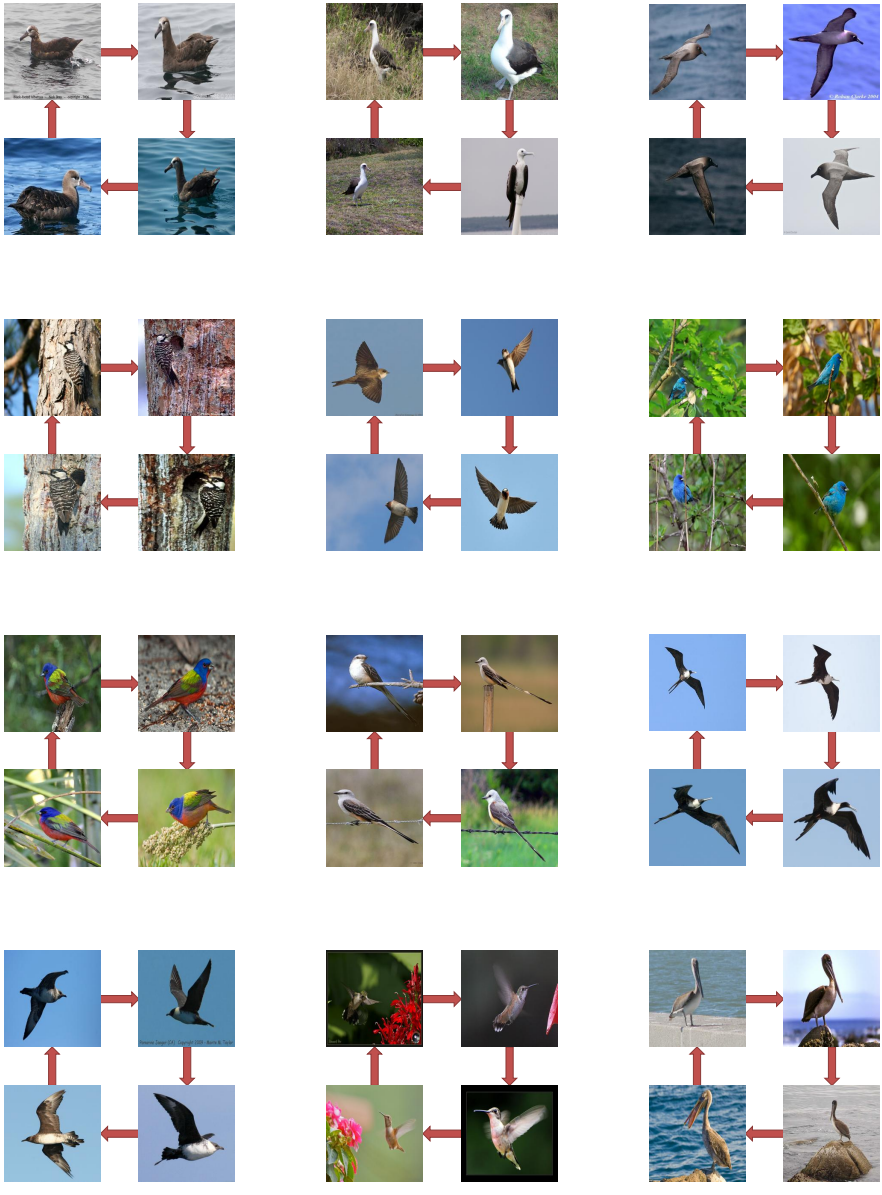


**Fig. 4.** Sample cycle detection results on the CUB-200-2011 dataset.

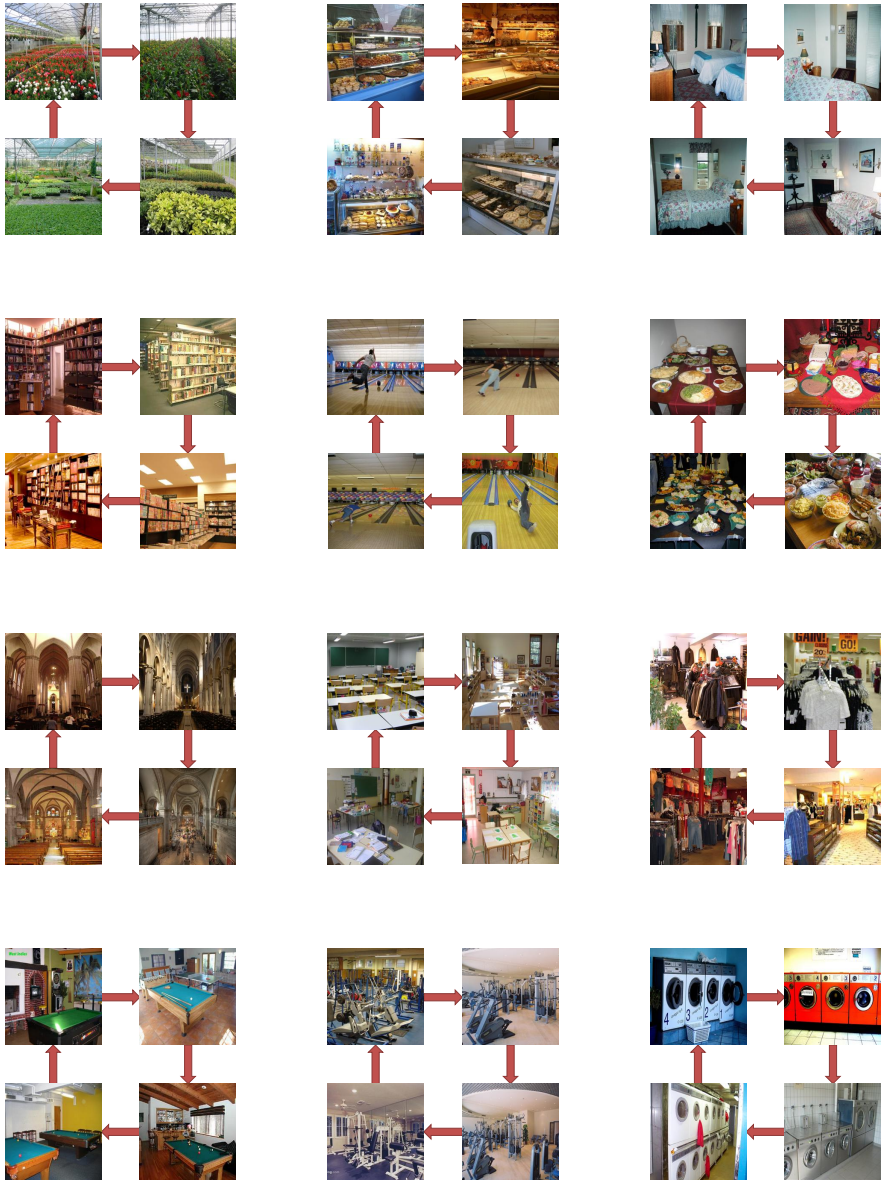
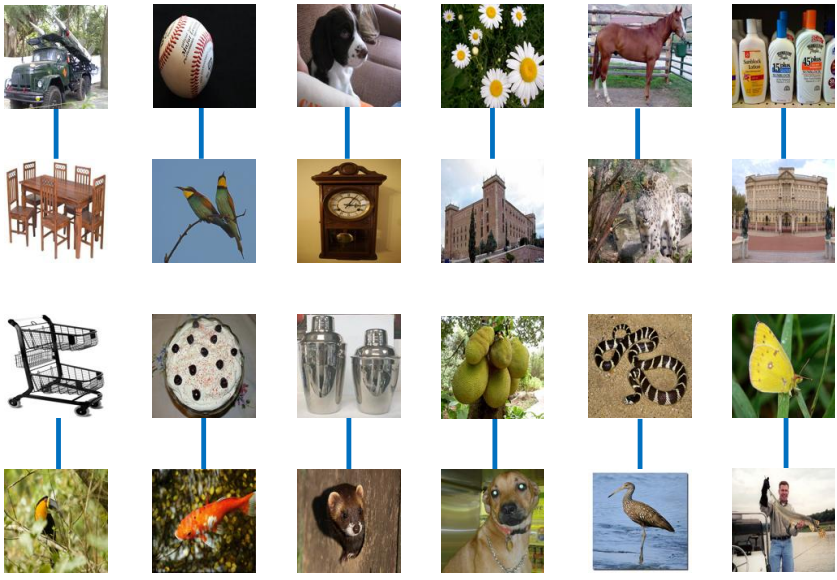
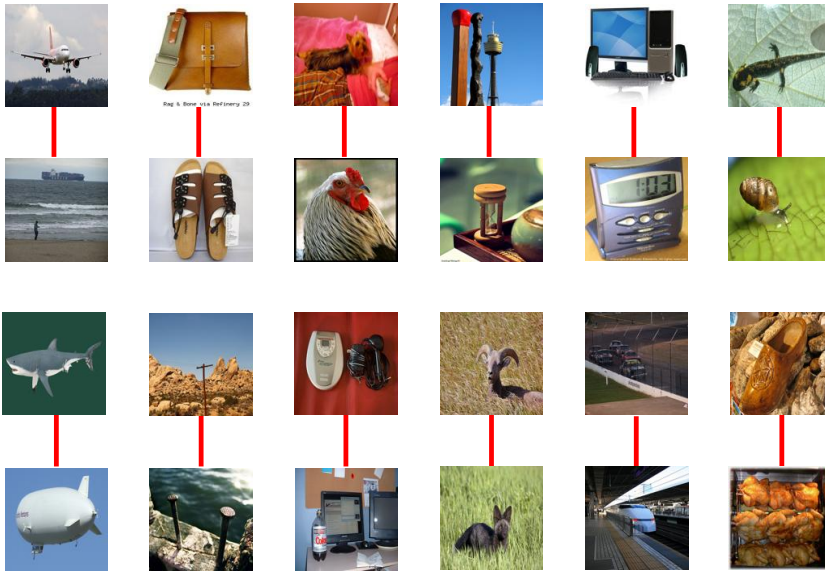


**Fig. 5.** Sample cycle detection results on the MIT indoor-67 dataset.



(a) Easy negative image pairs



(a) Hard negative image pairs

**Fig. 6.** Examples of negative image pairs. (a) Image pairs with large Euclidean distance have significant appearance differences. They are often easy samples which do not provide much information for learning a good CNN representation. (b) Image pairs with large geodesic distance are likely to belong to different visual categories but could be visually similar in appearances.

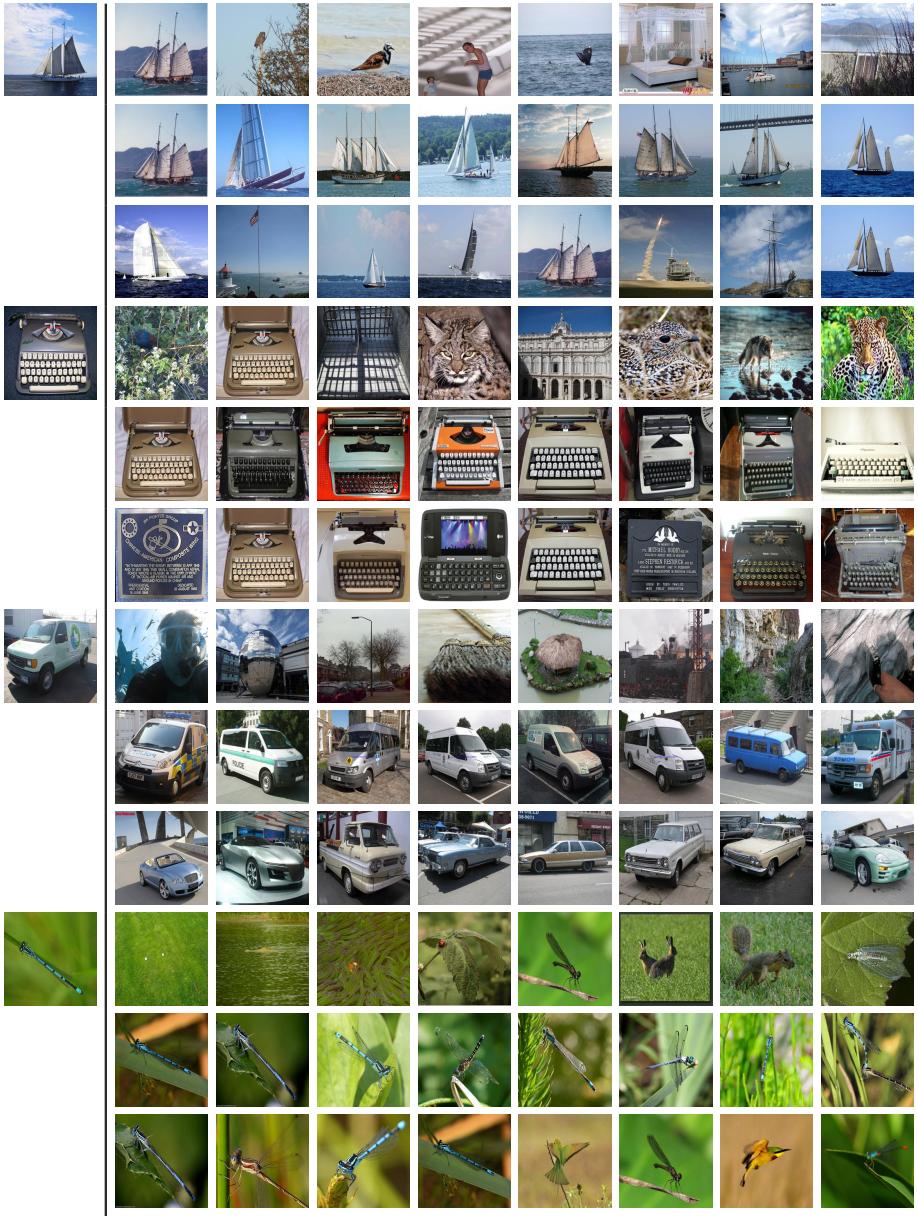


**Fig. 7.** Additional examples of nearest neighbor retrieval results. The query images are shown on the left hand side. For each query, the three rows show the top 8 nearest neighbors obtained by AlexNet with random parameters, AlexNet trained with full supervision and AlexNet trained using our unsupervised method, respectively. We use the FC7 features to compute Euclidean distance for all the three methods.

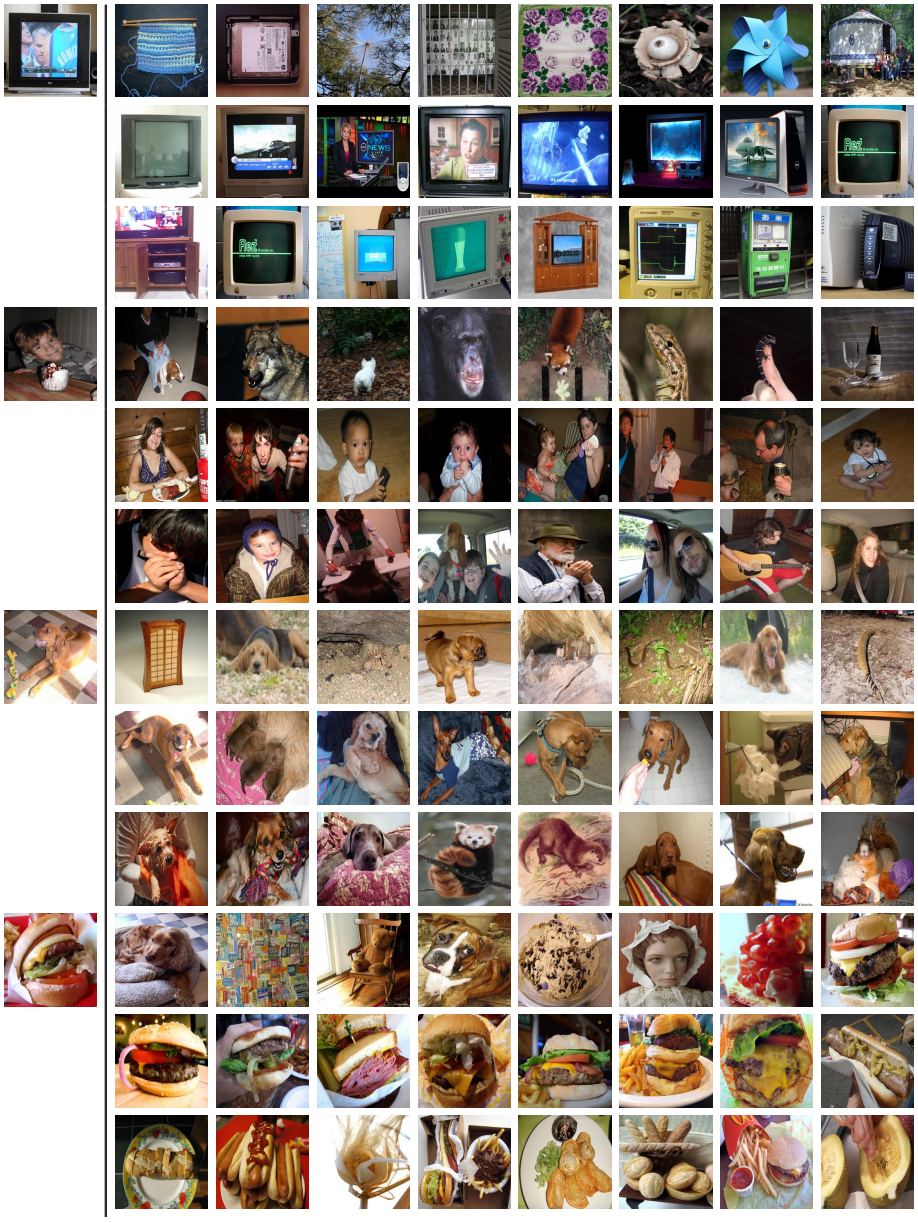


**Fig. 8.** Additional examples of nearest neighbor retrieval results. The query images are shown on the left hand side. For each query, the three rows show the top 8 nearest neighbors obtained by AlexNet with random parameters, AlexNet trained with full supervision and AlexNet trained using our unsupervised method, respectively. We use the FC7 features to compute Euclidean distance for all the three methods.

**Table 1.** Image classification performance using our unsupervised feature learning method in terms of average precision (AP) on the VOC 2007 *test* set.

| Methods | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Unsup. | 64.6 | 58.0 | 64.3 | 52.7 | 21.0 | 48.4 | 73.4 | 45.2 | 44.7 | 38.4 | |
| Unsup. + FT | 73.5 | 66.1 | 62.5 | 59.4 | 17.4 | 49.8 | 76.1 | 53.0 | 44.6 | 39.6 | |

| Methods | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Unsup. | 43.6 | 48.1 | 68.0 | 52.4 | 87.1 | 32.8 | 45.1 | 39.7 | 60.0 | 48.1 | 51.8 |
| Unsup. + FT | 53.4 | 50.7 | 74.1 | 67.9 | 84.7 | 41.4 | 48.8 | 40.9 | 72.5 | 54.5 | 56.5 |

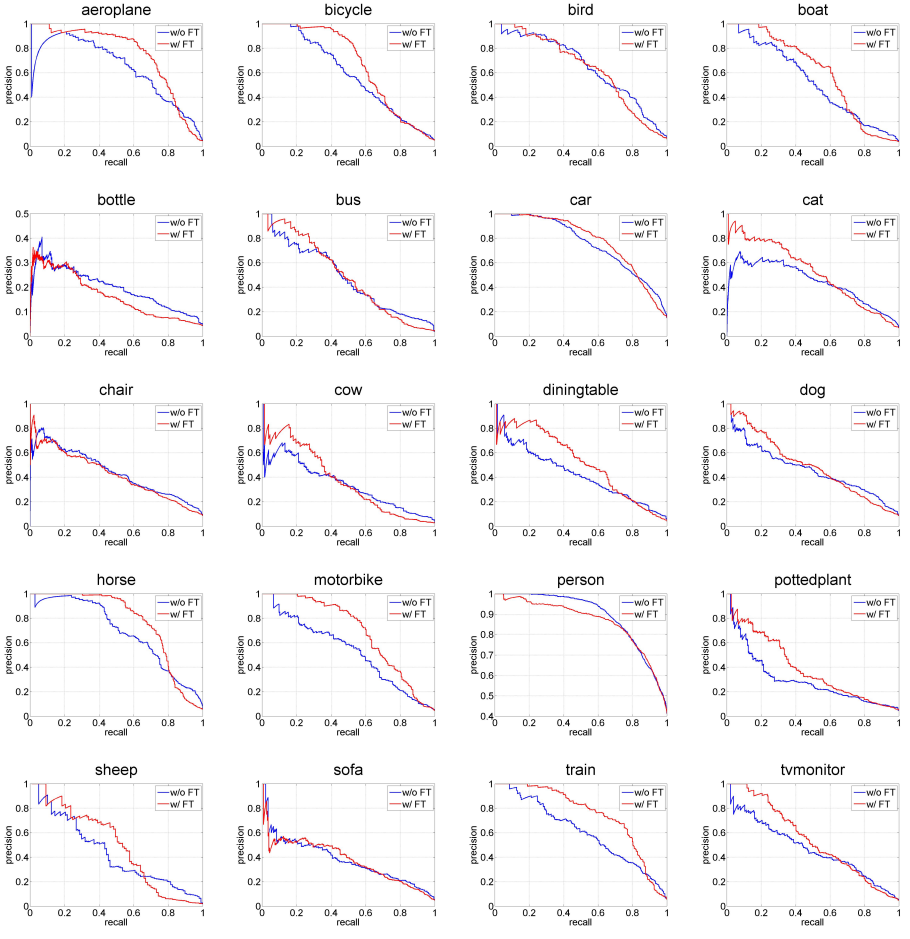



**Fig. 9.** Precision-recall curves for each category on the VOC 2007 *test* set.

**Table 2.** Mean classification accuracy on the CIFAR-10 dataset when $m$ images per class are annotated.

| | m=1 | m=5 | m=10 | m=50 | m=100 | m=500 |
|---|---|---|---|---|---|---|
| Baseline | 26.6 | 40.7 | 56.1 | 69.5 | 77.4 | 84.1 |
| Ours | 34.1 | 53.8 | 65.4 | 73.7 | 79.2 | 84.5 |

**Table 3.** Mean classification accuracy on the CUB-200-2011 dataset when $m$ images per class are annotated.

| | m=1 | m=5 | m=10 | m=20 | m=30 |
|---|---|---|---|---|---|
| Baseline | 11.2 | 26.3 | 36.3 | 48.7 | 52.5 |
| Ours | 14.7 | 27.7 | 39.0 | 48.6 | 53.1 |

**Table 4.** Mean classification accuracy on the MIT indoor-67 dataset when $m$ images per class are annotated.

| | m=1 | m=5 | m=10 | m=50 | m=80 |
|---|---|---|---|---|---|
| Baseline | 16.3 | 32.0 | 39.6 | 56.7 | 59.8 |
| Ours | 19.5 | 33.5 | 43.5 | 57.4 | 60.6 |