# Unsupervised Visual Representation Learning by Graph-based Consistent Constraints

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

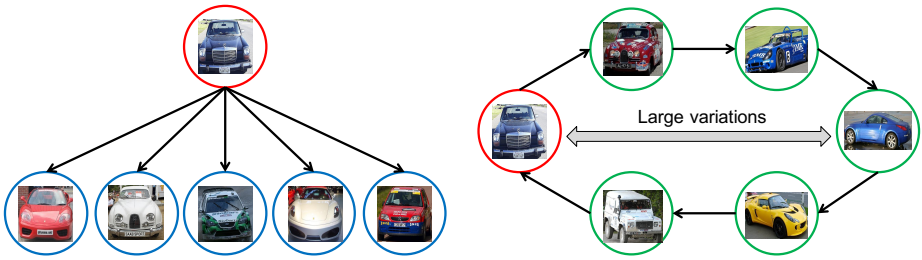
**Abstract.** Learning rich visual representations often require training on datasets of millions of manually annotated examples. This substantially limits the scalability of learning effective representations as labeled data is expensive or scarce. In this paper, we address the problem of unsupervised visual representation learning from a large, unlabeled collection of images. By representing each image as a node and each nearest-neighbor matching pair as an edge, our key idea is to leverage graph-based analysis to discover positive and negative image pairs (i.e., pairs belonging to the same and different visual categories). Specifically, we propose to use a cycle consistency criterion for mining positive pairs and geodesic distance in the graph for hard negative mining. We show that the mined positive and negative image pairs can provide accurate supervisory signals for learning effective representations using Convolutional Neural Networks (CNNs). We demonstrate the effectiveness of the proposed unsupervised constraint mining method in two settings: (1) unsupervised feature learning and (2) semi-supervised learning. For unsupervised feature learning, we obtain competitive performance with several state-of-the-art approaches on the PASCAL VOC 2007 dataset. For semi-supervised learning, we show boosted performance by incorporating the mined constraints on three image classification datasets.

**Keywords:** Unsupervised feature learning, semi-supervised learning, image classification, convolutional neural networks

## 1 Introduction

Convolutional neural networks have recently achieved impressive performance on a broad range of visual recognition tasks [1,2,3]. However, the success of CNNs is mainly attributed to supervised learning over massive amounts of human-labeled data. The need of large-scale manual annotations substantially limits the scalability of learning effective representations as labeled data is expensive or scarce. In this paper, we address the problem of *unsupervised* visual representation learning. Given only a large, unlabeled image collection, we aim to learn

⋆ Corresponding author.




(a) Direct matching (b) Cyclic matching

**Fig. 1.** Illustration of positive mining based on cycle consistency. (a) Direct image matching using similarity of the appearance features often results in matching pairs with very similar appearances (e.g., certain pose of cars). (b) By finding cycles in the graph, we observe that image pairs in the cycle are likely to belong to the same visual category but with large appearance variations (e.g., under different viewpoints).

rich visual representations without using any manual supervision. This particular setting is important for many practical applications because large amounts of interconnected visual data is readily available on the Internet. However, it remains challenging to learn effective representations for visual recognition in an unsupervised fashion.

Numerous efforts have been made on unsupervised learning [4,5,6,7,8]. Existing approaches aim to use reconstruction as a pretext task for visual representation learning. The most commonly used architecture is an autoencoder which aims at reconstructing input images from noisy ones [6,7,8]. However, current reconstruction-based algorithms tend to learn filters detecting low-level patterns (e.g., edges, textures). Such algorithms may not generalize well to high-level visual recognition tasks. Recent work explores various types of supervisory signals freely available in images and videos for unsupervised visual representation learning. Examples include ego-motion [9,10], context prediction [11], and tracking [12]. However, ego-motion information does not correlate with semantic information well. Spatial context prediction [11] and tracking [12] consider only *instance-level* data as the training samples are taken *within* the same image and video.

In this paper, we propose a new way to generate *category-level* training samples for unsupervised visual representation learning. The general idea is that we can discover underlying *semantic* similarity among images by leveraging graph-based analysis over a large collection of images. We construct the $k$-nearest neighbor ($k$-NN) graph by representing each image as a node and each nearest-neighbor matching pair as an edge. Unlike other methods that use the nearest neighbor graphs to learn similarity functions [13,14], we use the graph to mine constraints for learning rich visual representations. Specifically, we propose to use a cycle consistency criterion for mining positive pairs. Compared to the direct image matching, cycle consistency allows us to mine image pairs from the

same category yet with large appearance variations. The basic idea for positive mining is illustrated in Fig. 1. For negative image pair mining, we propose to use geodesic distance in the graph to discover hard negative samples. Image pairs with large geodesic distance are likely to belong to different categories but may have a small Euclidean distance in the original feature space. We observe that the mined positive and negative image pairs can provide accurate supervisory signals to train a CNN for learning effective representations. We validate the effectiveness of the proposed unsupervised constraint mining method in two settings: (1) unsupervised feature learning and (2) semi-supervised learning. For unsupervised feature learning, we obtain competitive performance with several state-of-the-art approaches on the PASCAL VOC 2007 dataset. For semi-supervised learning, we improve the classification results by incorporating the mined constraints on three datasets.

We make the following three contributions in this work:

1. We propose a simple but effective approach to mine semantically similar and dissimilar image pairs from a large, unlabeled collection of images.
2. We tackle the problem of learning rich visual representations in an unsupervised manner. Using the mined image pairs, we train a Siamese network to perform binary classification (i.e., same or different categories). Using the CNN model trained on the large-scale ImageNet dataset without any labels, we obtain competitive performance with the state-of-the-art unsupervised learning approaches on the PASCAL VOC 2007 dataset.
3. We show how the unsupervised constraint mining approach can also be used in a semi-supervised learning problem. We improve the classification accuracy by incorporating the mined constraints, particularly when the number of available training samples is limited.

## 2 Related Work

**Visual representation learning.** Convolutional neural networks have achieved great success on various recognition tasks [1,2,3]. Typical CNN-based visual representation learning approaches rely on full supervision, i.e., images with manually annotated class labels. Recent research on visual representation learning has been explored in a weakly supervised [15,16,17,18], semi-supervised [19,20] and unsupervised [11,12] fashion. Various types of supervisory signals are exploited to train CNNs as the substitutes for class labels. For example, Agrawal et al. [9] and Jayaraman et al. [10] train CNNs by exploiting ego-motion information. Wang et al. [12] track image patches in a video to train the network with a ranking loss. Doersch et al. [11] extract pairs of patches from an image and train the network to predict their relative positions. Chen et al. [21] and Joulin et al. [22] utilize the available large-scale web resources for learning CNN representations. However, ego-motion information [9,10] does not correlate with semantic information well. Spatial context prediction [11] and tracking [12] consider only *instance-level* data as the training samples are taken *within* the same image and video. In contrast, we use graph-based analysis to generate *category-level*

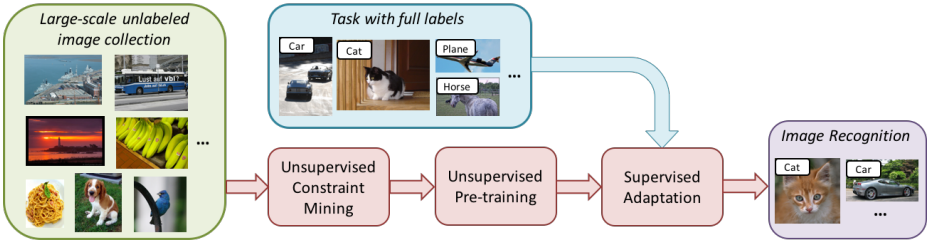
training samples *across* different images. These category-level samples contain positive image pairs that are semantically similar but may have large intra-class appearance variations. Such information is crucial for learning visual representations for factoring out nuisance appearance variations.

**Unsupervised object/patch mining.** Another line of related work is unsupervised object/patch mining. Existing methods use various forms of clustering or matching algorithms for object discovery [23], ROI detection [24] and patch mining [25,26]. Examples include spectral clustering [27], discriminative clustering [25,26], and alternating optimization algorithms [24,23]. However, clustering methods are typically sensitive to the pre-defined number of clusters. In contrast, our unsupervised constraint mining method aims at finding semantically similar and dissimilar image pairs instead of multiple image clusters. Compared to iterative optimization methods, our mining algorithm is more efficient and can be easily applied to large-scale datasets. In addition, we rely on CNNs to learn effective visual representations while most unsupervised object/patch mining methods use only hand-crafted features.

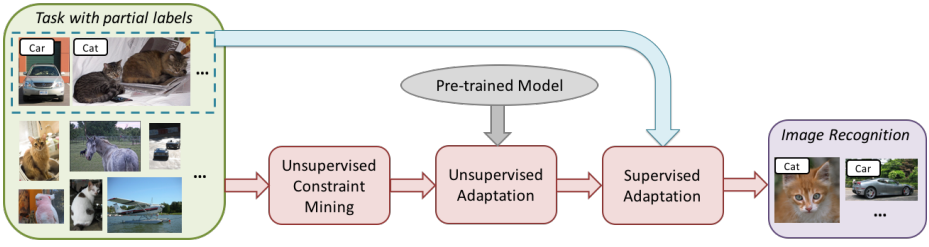
**Cycle consistency.** Cycle consistency in a graph has been applied to various computer vision and graphics problems, including co-segmentation [28,29], structure from motion [30,31] and image matching [32,33]. These approaches exploit cycles as constraints and solve constrained optimization problems for establishing correspondences among pixels/keypoints/patches across different images. In this work, we observe that cycle consistency can be used for finding semantically similar images. With detecting cycles in the $k$-NN graph, we can mine positive image pairs with large appearance variations from an unlabeled image collection. Our work is also related to symmetric nearest neighbor matching [34,35]. For example, Dekel et al. [35] match pairs of points where each point is the nearest neighbor of the other. This is a particular case (i.e., 2-cycle) of cycle consistency in our setup.

## 3 Overview

Our goal is to learn rich visual representations in an unsupervised manner. We propose an unsupervised constraint mining algorithm to generate category-level image pairs from an unlabeled image collection (Section 4). For positive pair mining, we detect cycles in the $k$-NN graph and take all the matching pairs in the cycles as positive samples. Compared to the direct image matching, image pairs mined by cycle consistency are likely to belong to the same visual category but with large appearance variations. For negative pair mining, we take image pairs with large geodesic distance in the graph as negative samples. Such mined negative pairs are likely to belong to different categories but may have a small Euclidean distance in the original feature space. We validate the effectiveness of the proposed unsupervised constraint mining algorithm in two settings: unsupervised feature learning (Section 5.1) and semi-supervised learning (Section 5.2). Fig. 2 shows the overview of the two settings.




(a) Unsupervised feature learning




(b) Semi-supervised learning

**Fig. 2.** Overview of the two settings for visual representation learning. For unsupervised feature learning, our goal is learning visual representations from a large-scale unlabeled image collection and employing the learned representations for specific recognition tasks with full labels. For semi-supervised learning, our goal is adapting visual representations from the supervised pre-trained model to specific recognition tasks with partial annotations.

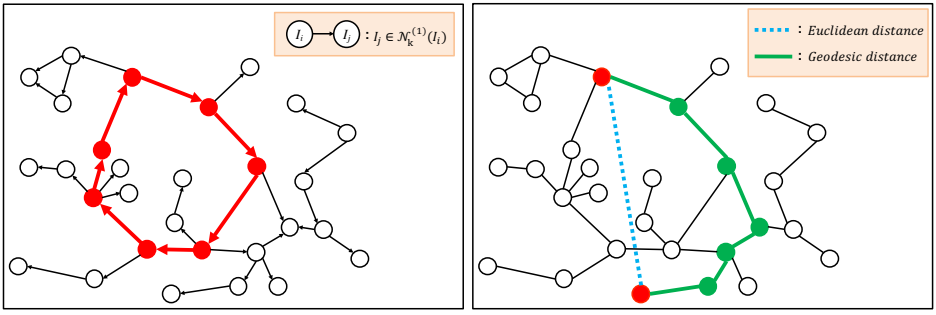
## 4 Unsupervised Constraint Mining

In this section, we introduce the unsupervised constraint mining algorithm. We start with computing the Euclidean distance between each image pair in the original feature space. We then construct a $k$-NN graph $G = (V, E)$. Each node $v \in V = \{I_1, I_2, \ldots, I_N\}$ denotes an image. Each directed edge $e_{ij}$ denotes a matching pair "$I_i \rightarrow I_j$" if $I_j$ belongs to the $k$-nearest neighbors of $I_i$. The edge weight $w_{ij}$ is defined by the Euclidean distance between the matching pair.

### 4.1 Positive constraint mining

We define that $I_j$ is an $n$-order $k$-nearest neighbor of the image $I_i$ if there exists a directed path of length $n$ from image $I_i$ to image $I_j$. The set of $n$-order $k$-nearest neighbors for image $I_i$ is denoted as $\mathcal{N}_k^{(n)}(I_i)$. For example, if $I_b$ belongs to the 5-nearest neighbors of $I_a$ and $I_c$ belongs to the 5-nearest neighbors of $I_b$, we have $I_c \in \mathcal{N}_5^{(2)}(I_a)$. Naturally, if $I_i$ belongs to its own $n$-order $k$-nearest neighbors, we then obtain a directed cycle.

$$I_i \in \mathcal{N}_k^{(n)}(I_i), \;\; n = 2, 3, 4, \ldots \; . \tag{1}$$




(a) Positive constraint mining (b) Negative constraint mining

**Fig. 3.** Illustration of the graph-based unsupervised constraint mining algorithm. (a) For positive mining, we propose to use cycle consistency to mine image pairs from the same class but with large appearance variations. (b) For negative mining, we propose to use geodesic distance to mine image pairs from the different classes but with a relatively small Euclidean distance in the original feature space.

For each node in the $k$-NN graph, we search its $n$-order $k$-nearest neighbors and detect cycles according to (1). An $n$-cycle constraint can generate $n(n-1)/2$ different pairs of images. We take these pairs as positive samples for the subsequent CNN training. Fig. 3(a) illustrates the process for positive constraint mining.

Cycle consistency offers two advantages for generating positive image pairs. (1) It helps mine indirect matching pairs from the same category yet with large appearance variations. For example, a 4-cycle constraint "$I_a \rightarrow I_b \rightarrow I_c \rightarrow I_d \rightarrow I_a$" will generate two indirect pairs as $(I_a, I_c)$ and $(I_b, I_d)$. Although the image $I_a$ and $I_c$ may have dramatically different appearances, the third image $I_b$ (or $I_d$) provides indirect evidence supporting their match. (2) It filters the large candidate set of $k$-NN matching pairs and selects the most representative ones (e.g., the adjacent pair $(I_a, I_b)$ in the 4-cycle constraint).

### 4.2 Negative constraint mining

Geodesic distance is widely used in manifold learning and has recently been applied to foreground/background segmentation as a low-level metric [36,37,38]. In our method, we use geodesic distance to mine hard negative image pairs in a $k$-NN graph. Specifically, we first use the Floyd-Warshall algorithm [39] for finding shortest paths between each node in the graph. The geodesic distance $g_{ij}$ is the accumulated edge weights along the shortest path from $I_i$ to $I_j$. We then perform random selection among those image pairs with large geodesic distance as negative samples. Fig. 3(b) illustrates the process for negative constraint mining.

Geodesic distance brings two advantages for generating negative image pairs. (1) Image pairs with large Euclidean distance are often easy samples, which do

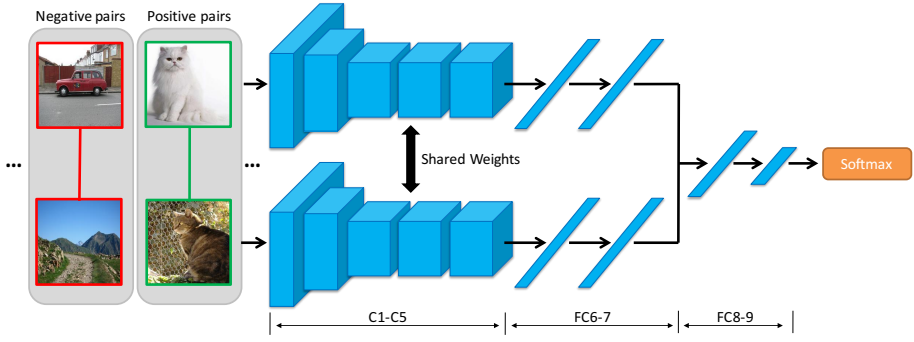



**Fig. 4.** The proposed Siamese network for binary classification. C1-FC7 layers follow the AlexNet architecture and share weights. FC8-9 layers have 64 and 2 neurons, respectively. A binary softmax classifier is used to predict whether the two images belong to the same category.

not contain much information for learning a good CNN representation. This is because the original Euclidean distance only expresses the appearance similarity between two images. In contrast, image pairs with large geodesic distance are likely to belong to different categories but may have small Euclidean distances in the original feature space. (2) Within a typical multi-class image dataset (e.g., the 1,000 classes in the ImageNet classification task), an overwhelming majority of random image pairs are negative samples. It is thus more efficient to select hard negative pairs based on geodesic distance for learning effective representations than collecting large amounts of easy samples.

## 5 Visual Representation Learning

To learn visual representations by the mined positive and negative pairs, we design a two-branch Siamese network for binary pair classification. Fig. 4 shows the Siamese network architecture. In our experiments, we take two images with size $227 \times 227$ as input. The layers of C1-FC7 follow the AlexNet architecture and share weights. We concatenate the two FC7 outputs and stack two fully connected layers of FC8-9 with 64 and 2 neurons, respectively. A softmax loss function is used to train the entire network for predicting whether the two images belong to the same category.

### 5.1 Unsupervised feature learning

In the setting of unsupervised feature learning (Fig. 2(a)), the goal is learning visual representations from a large-scale unlabeled image collection and employing the learned representations for specific recognition tasks with full labels. To this end, we first use the proposed unsupervised constraint mining algorithm to discover positive and negative pairs from the ImageNet 2012 dataset [40]

without any labels. We use Fisher Vectors based on dense SIFT [41] as feature descriptors.[1] Instead of directly applying our algorithm to the entire large-scale dataset with 1.2 million nodes, we randomly divide the training set into multiple subsets. Image pairs are mined in each subset and assembled eventually. In the unsupervised pre-training stage, we use the mined pairs to train the Siamese network (Fig. 4) for binary pair classification. Mini-batch Stochastic Gradient Descent (SGD) is used to train the network with random initialization. Section 6.1 describes more training details. In the supervised adaptation stage, we use the ground-truth data to fine-tune the network with a softmax loss for image classification.

### 5.2 Semi-supervised learning

In the setting of semi-supervised learning (Fig. 2(b)), the goal is adapting visual representations from the supervised pre-trained model to specific recognition tasks with partial annotations. We first use the proposed unsupervised constraint mining algorithm to mine positive and negative image pairs on the entire dataset. In the unsupervised adaptation stage, we use the mined pairs to train the Siamese network (Fig. 4), which is initialized using the pre-trained parameters on ImageNet with class labels. In the supervised adaptation stage, we use the partial ground-truth data to fine-tune the base network with the softmax loss for image classification.

## 6 Experiments

### 6.1 Implementation details

We use Caffe [42] to train our network with a Tesla K40 GPU. In all experiments, SGD is used for optimization with the batch size of 50. Each batch contains 25 positive pairs and 25 negative pairs.

For unsupervised feature learning, we randomly divide the entire ImageNet training set into 128 subsets where each subset contains ∼10k images. In total, our method mines ∼1 million positive pairs and ∼13 million negative pairs. We train the network from random initialization with 400k iterations. The learning rate is initially set to 0.01 and follows a polynomial decay with the power parameter of 0.5. It takes six hours to mine the pairs and five days to train the network.

For semi-supervised learning, we use the unsupervised mined pairs to train the Siamese network with the fixed learning rate of 0.001 for 50k iterations. In the supervised adaption stage, all available image labels are used to fine-tune the base network with the fixed learning rate of 0.001 for 5k iterations.

The source code, as well as the pre-trained models, is available at the project webpage.

[1] For efficiency, PCA is used to project the high-dimensional FV descriptors to 512 dimensions.

### 6.2 Datasets and evaluation metrics

We evaluate the image classification performance of the unsupervised learned representations on the PASCAL VOC 2007 dateset [43]. The challenging PASCAL VOC dataset contains 20 objects categories with large intra-class variations in complex scenes. We use three datasets to evaluate the recognition performance of semi-supervised learning: (1) CIFAR-10 for object recognition [44], (2) CUB-200-2011 for fine-grained recognition [45] and (3) MIT indoor-67 for scene recognition [46]. We use average precision (AP) as the metric for image classification on VOC 2007 and top-1 classification accuracy for the other three datasets.

### 6.3 Controlled experiments

**Evaluation on positive mining.** We compare the proposed positive mining method with random sampling and direct matching for image classification on CIFAR-10. For fair comparisons, we randomly sample the same set of 500k true negative pairs for the three positive mining methods.[2]

- Random sampling: Randomly sampling 10k pairs.
- Direct matching: The top 10k pairs with the smallest Euclidean distance.
- Cycle consistency: The 10k pairs mined by $n$-cycle constraints with $k$=4.

We use the positive pairs mined by different methods (along with the same negative pairs) to train the Siamese network. We initialize the base network using the pre-trained parameters on ImageNet with class labels. For testing, we extract 4096-d FC7 features and train linear SVMs for classification. Table 1 shows the mining and classification results with different positive mining methods. In terms of true positive rate, cycle consistency significantly outperforms random sampling and direct matching. The results demonstrate that our method can handle large intra-class variations and discover accurate pairs from the same category. Regarding the classification accuracy, using 4-cycle constraints achieves significant improvement over direct similarity matching by around 3 points. The experimental results demonstrate that cycle consistency helps learn better CNN feature representations. We also observe that the recognition performance is insensitive to the cycle length, which shows the stability and robustness of the proposed method. Notably, although 2-cycle and 3-cycle constraints do not generate indirect matching pairs, they are crucial for selecting representative positive pairs for feature learning. Without cycle consistency, acyclic transitive matching easily generates false positive pairs, particularly when the cycle length $n$ is large. We believe that cycle consistency provides an effective criterion to discover good positive pairs for learning effective representations.

**Parameter analysis.** Fig. 5 shows the statistics of mined cycles with different $k$ (the number of nearest neighbors) and $n$ (the length of cycle). The amount of mined cycles increases as $k$ increases because larger $k$ results in more linked nodes

[2] True positives (TP) and true negatives (TN) are denoted as those pairs belonging to the same and different visual categories, respectively.

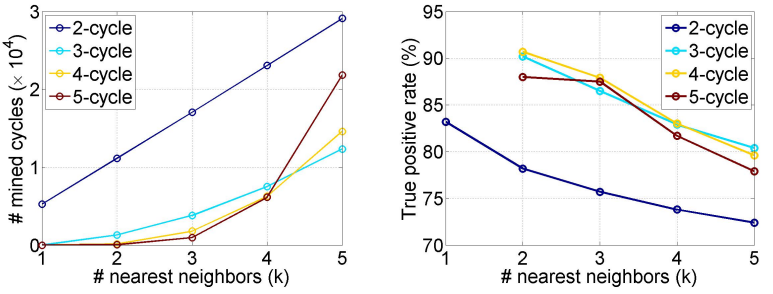



**Fig. 5.** The statistics of the mined cycle constraints on the CIFAR-10 *train* set. **Left:** Total amount of mined cycles. **Right:** True positive rate among all the mined pairs.

**Table 1.** Comparisons of different positive mining methods on CIFAR-10.

| | Random sampling | Direct matching | 2-cycle | 3-cycle | 4-cycle | 5-cycle |
|---|---|---|---|---|---|---|
| TP rate | 10.0 | 59.0 | 73.8 | 82.9 | **83.0** | 81.7 |
| Accuracy | 73.7 | 78.0 | 79.9 | 80.5 | **80.9** | 80.2 |

**Table 2.** Comparisons of different negative mining methods on CIFAR-10.

| | Random sampling | Original distance | Geodesic distance |
|---|---|---|---|
| TN rate | 90.0 | **95.5** | 91.0 |
| Accuracy | 83.8 | 68.3 | **85.2** |

in the graph. On the other hand, as $k$ increases, the true positives rate drops due to the noise introduced by nearest neighbor matching. However, using 4-cycle constraints, we obtain a much higher true positive rate with a 40% relative improvement over direct matching (see Table 1). The results show that cycles do help get rid of the noise in the matching process.

**Effect of different features.** We evaluate different features for constructing the graph and obtain similar classification performance on CIFAR-10 (LBP: 76.7%, HOG: 80.7%, and SIFT+FV: 80.9%). The results show that cycle consistency works well on different hand-crafted features. We also use the initial ImageNet-pretrained CNN features to construct the graph. It achieves 81.6% accuracy on CIFAR-10, slightly higher than that of using SIFT+FV (80.9%).

**Evaluation on negative mining.** We conduct controlled experiments to examine the effectiveness of the proposed negative mining method on CIFAR-10. The same 500k true positive pairs are randomly sampled for the following three methods.

- Random sampling: Randomly sampling 500k pairs.
- Original distance: The top 500k pairs with the largest Euclidean distance.
- Geodesic distance: The 500k pairs mined with geodesic distance.

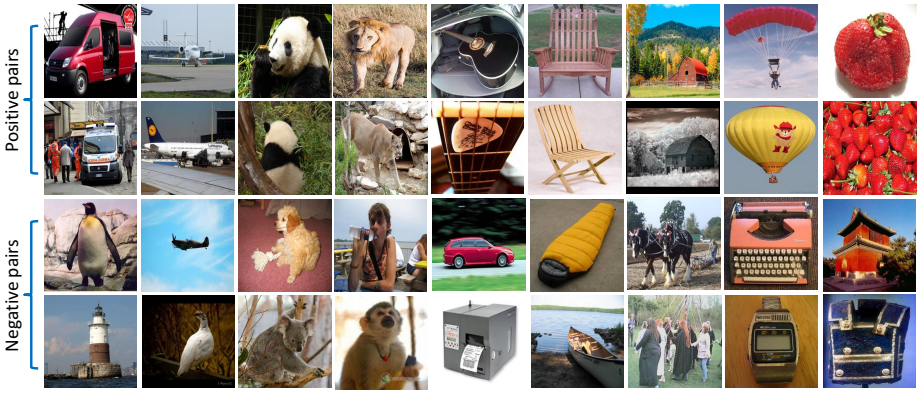



**Fig. 6.** Examples of positive and negative image pairs mined from the ImageNet 2012 dataset using our unsupervised constraint mining method.

We use the negative pairs mined by different methods (along with the same positive pairs) to train the Siamese network. Table 2 shows the mining and classification results with the three negative mining methods on CIFAR-10. The graph-based geodesic distance achieves classification accuracy of 85.2%, significantly outperforming the method by the original Euclidean distance by 17 points. Although more accurate pairs are mined by the original distance, they are often easy negative samples and do not provide much information for learning effective representations. Negative mining by random sampling performs well because an overwhelming majority of image pairs are negative in a typical image dataset, e.g., 90% on CIFAR-10. In general, the experimental results demonstrate that the proposed graph-based geodesic distance can generate hard negative samples to learn better representations for visual recognition.

### 6.4 Unsupervised learning results

**Qualitative evaluation.** We first show qualitative results obtained by our unsupervised feature learning method. Fig. 6 shows some examples of image pairs mined from the ImageNet 2012 dataset using the proposed unsupervised constraint mining method. Cycle consistency can mine positive pairs with large appearance variations (e.g., different viewpoints and shape deformations). Geodesic distance can mine hard negative pairs which share appearance similarities to an extent (e.g., bird and aeroplane, monkey and human). Fig. 7 shows examples of nearest neighbor search results using different feature representations. Our unsupervised method obtains similar retrieval results with the supervised pre-trained AlexNet for different types of visual categories.

**Quantitative evaluation.** We compare the proposed unsupervised feature learning method with several state-of-the-art approaches for image classification on VOC 2007 in Table 3. All the results are obtained by fine-tuning using

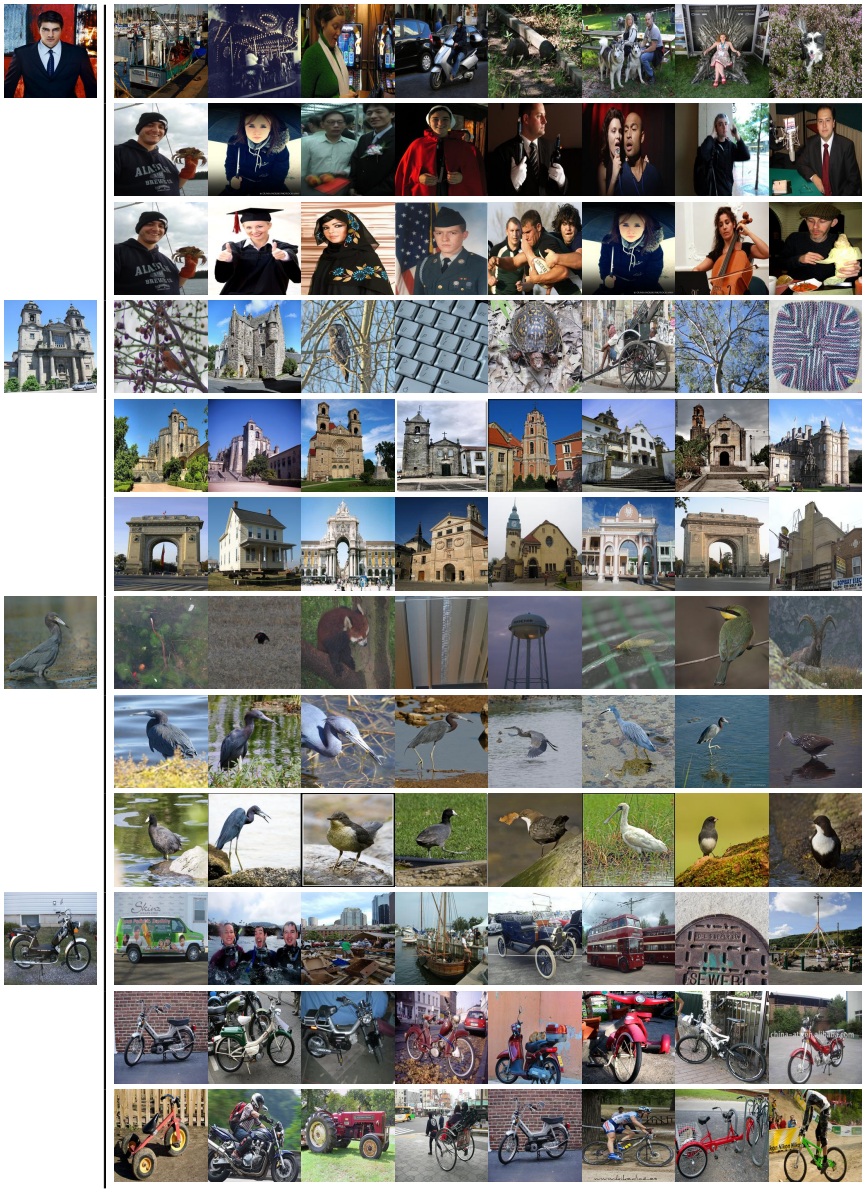


**Fig. 7.** Examples of nearest neighbor search results. The query images are shown on the far left. For each query, the three rows show the top 8 nearest neighbors obtained by AlexNet with random parameters, AlexNet trained with full supervision, and AlexNet trained using our unsupervised method, respectively.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

**Table 3.** Comparisons of classification performance on the VOC 2007 *test* set.

| Methods | Supervision | Classification |
|---|---|---|
| Agrawal et al. [9] | Ego-motion | 52.9 |
| Doersch et al. [11] | Context | 55.3 |
| Wang et al. [12] | Tracking triplet | **58.4** |
| Ours | Matching pair | 56.5 |
| Krizhevsky et al. [1] | Class labels | **69.5** |

the VOC 2007 training data.[3] We achieve competitive performance with the state-of-the-art unsupervised approaches. Compared to Agrawal et al. [9], we show a significant performance gain by 3.6 points. Ego-motion information does not correlate well with semantic similarity, and hence the trained model does not perform well for visual recognition. Our method outperforms Doersch et al. [11] which use context prediction as supervision. They consider only instance-level training samples within the same image while we mine category-level samples across different images. Wang et al. [12] achieve better performance by leveraging visual tracking of video data. However, our method aims at mining matching pairs from an unlabeled image collection. For fair comparisons, we use random initialization as in existing unsupervised feature learning work and do not include other initialization strategies.

We compare the classification performance using SIFT+FV and our learned features. Our learned features significantly outperform SIFT+FV by 10.5 points (56.5% vs. 46.0%). The results show that we do not train the network to replicate hand-crafted features. While we use hand-crafted features to construct the graph, the proposed graph-based analysis can discover underlying semantic similarity among unlabeled images for learning effective representations.

**Effect of network architectures.** We also evaluate the performance using GoogLeNet as the base network. We achieve 56.6% mAP on VOC 2007, which is similar with that of using AlexNet (56.5%).[4]

### 6.5 Semi-supervised learning results

We also evaluate the proposed unsupervised constraint mining algorithm in the semi-supervised setting. For the three datasets used, we randomly select several images per class on the training set as the partial annotated data. Fig. 8(a) shows that we achieve significant performance gains compared with directly fine-tuning on CIFAR-10. In the extreme case that only one image label per class is known, our method largely improves the mean accuracy by 7.5 points (34.1% vs. 26.6%). Using 4,000 labels of CIFAR-10, our method outperforms Rasmus et al. [48] from 79.6% to 84.3%. The experimental results demonstrate that our unsupervised

[3] The baseline numbers of [9,11,12] are from [47].

[4] The GoogLeNet-based Siamese network is trained with a batch size of 32 and 960k iterations.

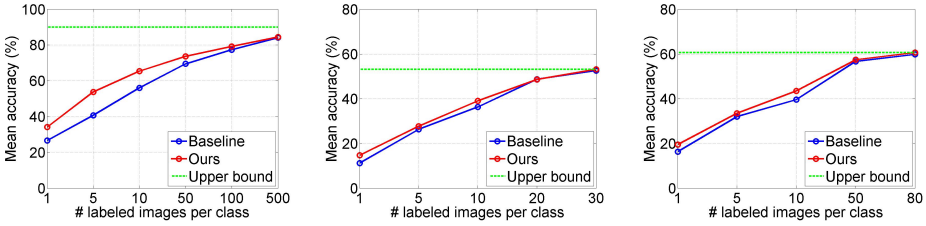



**Fig. 8.** Mean classification accuracy in the semi-supervised leaning tasks on three datasets: (a) CIFAR-10, (b) CUB-200-2011, and (c) MIT indoor-67. The upper bound represents the mean classification accuracy when images in the training set are fully annotated.

constraint mining method provides new useful constraints beyond annotations and helps better transfer the pre-trained network for visual recognition.

Fig. 8(b) and (c) show another two semi-supervised learning results on CUB-200-2011 and MIT indoor-67, respectively. The results show boosted classification performance for both fine-grained objects and scene categories. We obtain the true positive rate of 55.8% by 4-cycle constraints on CUB-200-2011 (only 0.5% by random sampling) and 65.8% on MIT indoor-67 (only 1.5% by random sampling). The results demonstrate that our method can generate accurate image pairs despite small inter-class differences among visual categories.

## 7 Conclusions

In this paper, we propose to leverage graph-based analysis to mine constraints from an unlabeled image collection for visual representation learning. We use a cycle consistency criterion to mine positive image pairs and geodesic distance to mine hard negative samples. The proposed unsupervised constraint mining method is applied to both unsupervised feature learning and semi-supervised learning. In the unsupervised setting, we mine a collection of image pairs from the large-scale ImageNet dataset without any labels for learning CNN representations. The learned features achieve competitive recognition results on VOC 2007 compared with existing unsupervised approaches. In the semi-supervised setting, we show boosted performance on three image classification datasets. In summary, our method provides new insights into data mining, unsupervised feature learning, and semi-supervised learning, and has broad applications for large-scale recognition tasks.

**Acknowledgments.** This work is supported in part by the Initiative Scientific Research Program of Ministry of Education under Grant #20141081253. J.-B. Huang and N. Ahuja are supported in part by Office of Naval Research under Grant N00014-16-1-2314. W.-C. Hung and M.-H. Yang are supported in part by the NSF CAREER Grant #1149783 and gifts from Adobe and Nvidia.

## References


1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS. (2012) 1, 3, 13
2. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014) 1, 3
3. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 1, 3
4. Le, Q.V.: Building high-level features using large scale unsupervised learning. In: ICML. (2012) 2
5. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: NIPS. (2012) 2
6. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: NIPS. (2007) 2
7. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: ICML. (2008) 2
8. Ranzato, M.A., Huang, F.J., Boureau, Y.L., LeCun, Y.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: CVPR. (2007) 2
9. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: ICCV. (2015) 2, 3, 13
10. Jayaraman, D., Grauman, K.: Learning image representations tied to ego-motion. In: ICCV. (2015) 2, 3
11. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV. (2015) 2, 3, 13
12. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV. (2015) 2, 3, 13
13. Malisiewicz, T., Efros, A.: Beyond categories: The visual memex model for reasoning about object relationships. In: NIPS. (2009) 2
14. Cao, S., Snavely, N.: Graph-based discriminative learning for location recognition. In: CVPR. (2013) 2
15. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?–weakly-supervised learning with convolutional neural networks. In: CVPR. (2015) 3
16. Li, D., Huang, J.B., Li, Y., Wang, S., Yang, M.H.: Weakly supervised object localization with progressive domain adaptation. In: CVPR. (2016) 3
17. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: CVPR. (2015) 3
18. Liang, X., Liu, S., Wei, Y., Liu, L., Lin, L., Yan, S.: Towards computational baby learning: A weakly-supervised approach for object detection. In: ICCV. (2015) 3
19. Chen, X., Shrivastava, A., Gupta, A.: NEIL: Extracting visual knowledge from web data. In: ICCV. (2013) 3
20. Divvala, S., Farhadi, A., Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning. In: CVPR. (2014) 3
21. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: ICCV. (2015) 3
22. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: ECCV. (2016) 3
23. Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: CVPR. (2015) 4

24. Kim, G., Torralba, A.: Unsupervised detection of regions of interest using iterative link analysis. In: NIPS. (2009) 4
25. Singh, S., Gupta, A., Efros, A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV. (2012) 4
26. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes paris look like paris? ACM Transactions on Graphics **31**(4) (2012) 4
27. Grauman, K., Darrell, T.: Unsupervised learning of categories from sets of partially matching image features. In: CVPR. (2006) 4
28. Wang, F., Huang, Q., Guibas, L.: Image co-segmentation via consistent functional maps. In: ICCV. (2013) 4
29. Wang, F., Huang, Q., Ovsjanikov, M., Guibas, L.: Unsupervised multi-class joint image segmentation. In: CVPR. (2014) 4
30. Wilson, K., Snavely, N.: Network principles for SfM: Disambiguating repeated structures with local context. In: ICCV. (2013) 4
31. Zach, C., Klopschitz, M., Pollefeys, M.: Disambiguating visual relations using loop constraints. In: CVPR. (2010) 4
32. Huang, Q.X., Guibas, L.: Consistent shape maps via semidefinite programming. In: SGP. (2013) 4
33. Zhou, T., Lee, Y.J., Yu, S.X., Efros, A.A.: FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In: CVPR. (2015) 4
34. Zhang, S., Yang, M., Cour, T., Yu, K., Metaxas, D.N.: Query specific fusion for image retrieval. In: ECCV. (2012) 4
35. Dekel, T., Oron, S., Rubinstein, M., Avidan, S., Freeman, W.T.: Best-buddies similarity for robust template matching. In: CVPR. (2015) 4
36. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: ECCV. (2012) 6
37. Krähenbühl, P., Koltun, V.: Geodesic object proposals. In: ECCV. (2014) 6
38. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: CVPR. (2015) 6
39. Floyd, R.W.: Algorithm 97: shortest path. Communications of the ACM **5**(6) (1962) 345 6
40. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. (2009) 7
41. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. IJCV **105**(3) (2013) 222–245 8
42. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM MM. (2014) 8
43. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV **88**(2) (2010) 303–338 9
44. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Computer Science Department, University of Toronto (2009) 9
45. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset. Technical report, California Institute of Technology (2011) 9
46. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR. (2009) 9
47. Krähenbühl, P., Doersch, C., Donahue, J., Darrell, T.: Data-dependent initializations of convolutional neural networks. In: ICLR. (2016) 13
48. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. In: NIPS. (2015) 13