# Instance Similarity Learning for Unsupervised Feature Representation

Ziwei Wang[1,2,3], Yunsong Wang[1], Ziyi Wu[1], Jiwen Lu[1,2,3]*, Jie Zhou[1,2,3]
[1] Department of Automation, Tsinghua University, China
[2] State Key Lab of Intelligent Technologies and Systems, China
[3] Beijing National Research Center for Information Science and Technology, China
{wang-zw18, wangys16}@mails.tsinghua.edu.cn; dazitu616@gmail.com;
{lujiwen,jzhou}@tsinghua.edu.cn

## Appendix A: Results on Object Detection

We first introduce the dataset that we carried out experiments on: The PASCAL VOC dataset consists of $9,963$ natural images from 20 different classes. Our model was trained on the PASCAL VOC 2007 trainval sets which contained $5,011$ images, and was evaluated on the PASCAL VOC 2007 test set including $4,952$ images. We utilized the mean average precision (mAP) as the evaluation metric. In our experiments, the two-stage detection framework Fast R-CNN [6] and Faster R-CNN [12] were applied with the AlexNet [9] and ResNet50 [8] as the backbone networks, which were pretrained on ImageNet with unlabeled data using our ISL method and then finetuned on PASCAL VOC 2007 for object detection.

We compare our method with the state-of-the-art unsupervised features including the clustering method DeepCluster[2], the instance specificity analysis methods Instance [13], MoCo-v1 [7] and MoCo-v2 [4] and the neighborhood discovery methods LA [14]. Table 1 shows the experimental results. We first trained the backbone by the listed unsupervised feature learning methods, and then loaded the weights as the pretrained model for finetuning the detection model on PASCAL VOC 2007. We also provide the performance of supervised pretraining for reference. Our ISL significantly outperforms the state-of-the-art neighborhood discovery method LA by a large margin, and is even comparable with the supervised methods when applying AlexNet as the backbone.

MoCo-v1 [7] validated that building large and consistent dictionary on-the-fly by momentum contrast enabled effective and efficient largescale contrastive learning, and SimCLR [3] verified that more data augmentation and an extra MLP head improved contrastive learning significantly. As a result, we combined our method with MoCo-v2 [4] which integrated techniques in MoCo-v1 and SimCLR to further enhance our ISL. We obtained the accuracy of MoCo-v2 by rerunning the code from the officially released code. Since our ISL provides informative supervision for contrastive learning which is obtained by neighborhood discovery based on geodesic distance of feature manifold, we further strengthen the feature discriminality when pretraining the detection model with our ISL on ImageNet. The pretrained features acquired by the integrated method MoCo-v2+ISL leads to significantly higher mAP on object detection compared with the supervisedly pretrained features, which demonstrates that the unsupervised feature learning methods is more effective to transfer the knowledge from ImageNet to PASCAL VOC 2007.

Table 1. Mean average precision (%) on PASCAL VOC 2007, where the architectures of AlexNet and ResNet50 were used as the backbone. Fast R-CNN and Faster R-CNN were applied as the detection framework.

| Method | AlexNet-Fast | AlexNet-Faster | ResNet50-Faster |
|---|---|---|---|
| Supervised | 56.8 | 54.3 | 74.6 |
| DeepCluster | 55.4 | – | – |
| Instance | 48.1 | 53.1 | 65.4 |
| LA | – | 53.5 | 69.1 |
| ISL | **56.6** | **54.2** | 70.0 |
| MoCo-v1 | – | – | 74.9 |
| MoCo-v2 | – | – | 76.3 |
| MoCo-v2+ISL | – | – | **77.6** |

## Appendix B: Results on Transfer Learning

In order to show the performance of our ISL on transfer learning, we conducted experiments to evaluate the features pretrained on ImageNet with different unsupervised learning methods. We employed the ResNet50 architectures for evaluation, and applied the CIFAR-10, CIFAR-100, Aircraft [10], Flowers [11], Food [1] and Caltech101 [5] datasets. Table 2 shows the conv5 feature accuracy of linear evaluation. MoCo-v2+ISL clearly outperforms the vanilla MoCo-v2 on all list datasets, which demonstrates the effectiveness of feature manifold mining in unsupervised learning.

*Corresponding author

Table 2. Classification accuracy (%) on CIFAR-10, CIFAR-100, Aircraft, Flowers, Food and Caltech101 datasets, where the ResNet50 architecture was used as the backbone.

| Method | CIFAR-10 | CIFAR-100 | Aircraft | Flowers | Food | Caltech101 |
|---|---|---|---|---|---|---|
| MoCo-v2 | 89.8 | 69.9 | 47.2 | 88.2 | 68.5 | 90.3 |
| MoCo-v2+ISL | **90.5** | **71.0** | **48.4** | **91.7** | **68.9** | **92.1** |

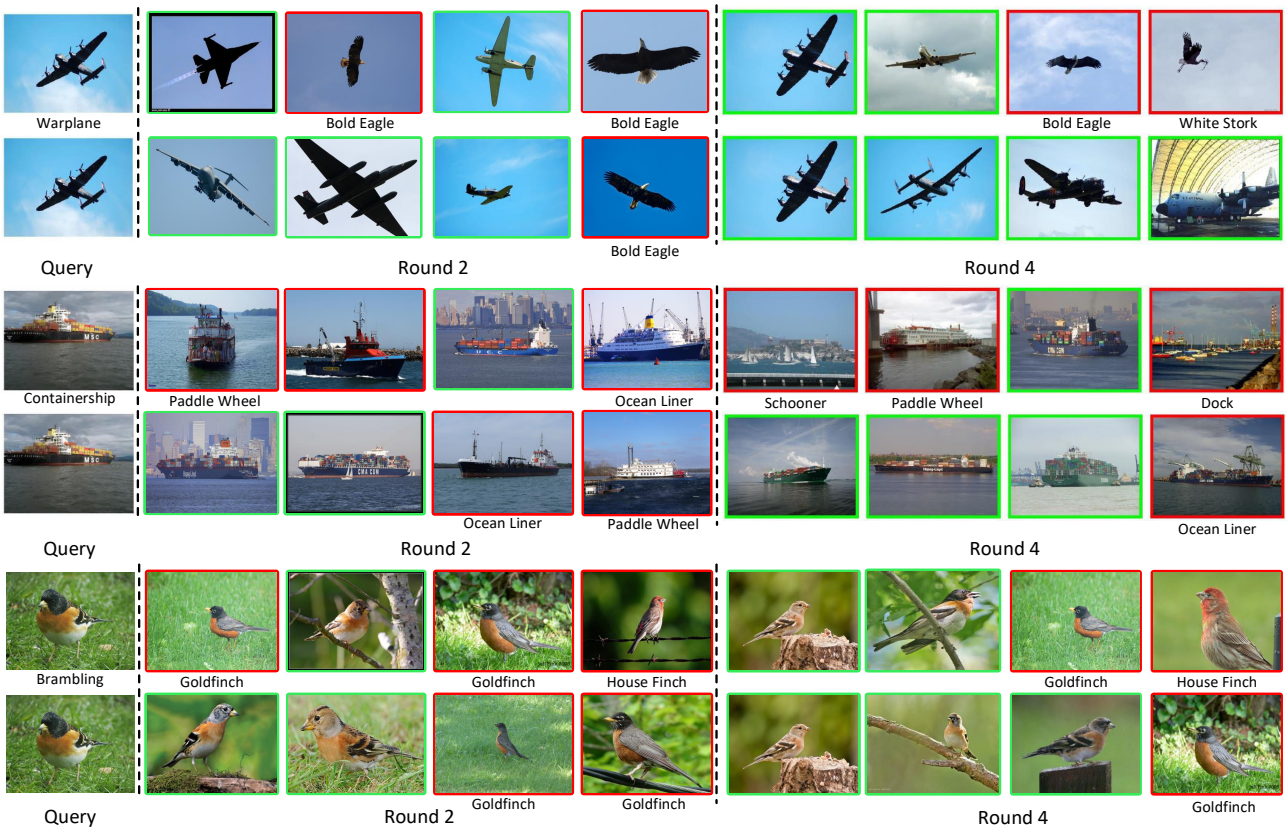


Figure 1. Examples of positive sample mining via LA (top row) and our ISL (bottom row) in different rounds during training. The images in green boxes represent the positives mined correctly and those in red boxes mean the images from other classes. We also offer the names of classes for anchors and the mistakenly mined positive samples below the images.

## Appendix C: Visualization of Positive Sample Mining

Figure 1 illustrates several examples of positive sample mining via LA and the proposed ISL in different rounds during training. For each anchor, images in the top and bottom row depict the mined positives via LA and ISL respectively. The green box stands for the instances that share the same label with the anchor while the red box means that the samples are in different classes with the anchor. We also offer the names of classes for anchors and the mistakenly mined positive samples below the images. The mined positive samples become more accurate with the increase of the rounds during training. However, LA regards instances with similar appearance including colors and shapes as positive samples and fails to distinguish the fine-grained difference among various classes. On the contrary, our ISL mines the feature manifold to assign similarity among instances according to the geodesic distance and successfully finds the semantically similar samples even with different appearance.

## References


[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014.

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pages 178–178, 2004.

[6] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[10] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[11] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[13] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[14] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, pages 6002–6012, 2019.