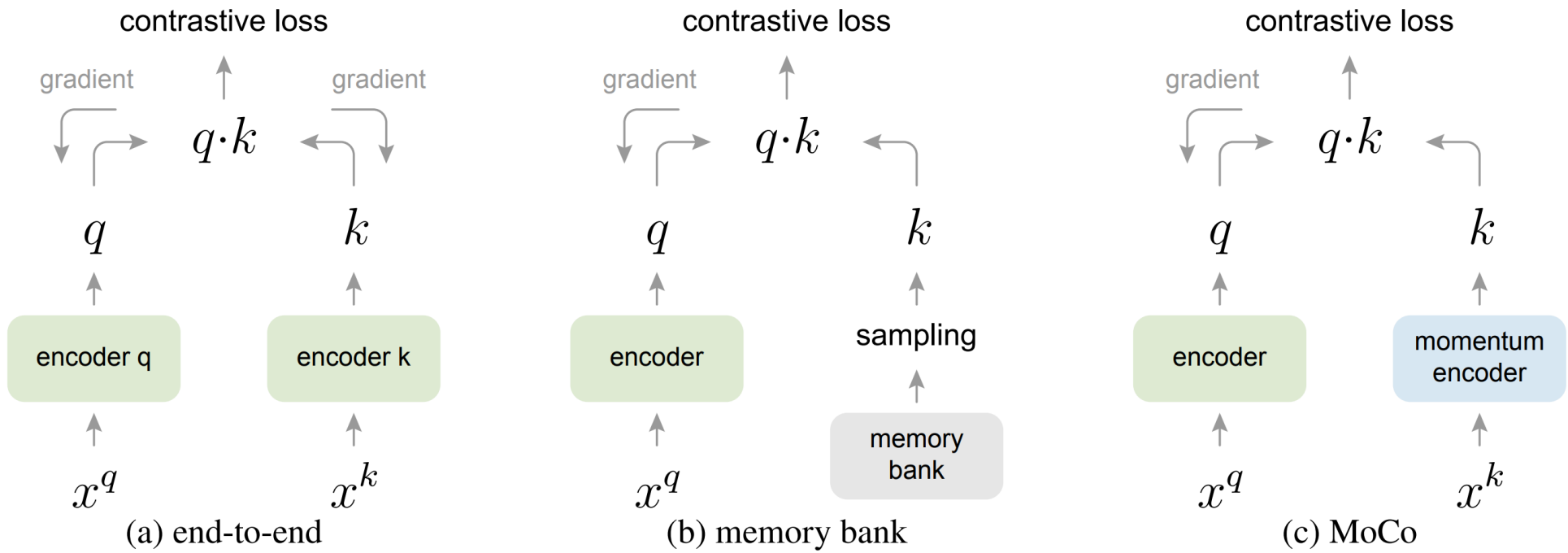

# Unsupervised Representation Learning by Invariance Propagation

**Feng Wang, Huaping Liu,* Di Guo, Fuchun Sun**
Department of Computer Science and Technology, Tsinghua University, China
Beijing National Research Center for Information Science and Technology
wang-f20@mails.tsinghua.edu.cn,hpliu@tsinghua.edu.cn
guodi.gd@gmail.com,fcsun@tsinghua.edu.cn

NIPS-2020

# Motivation

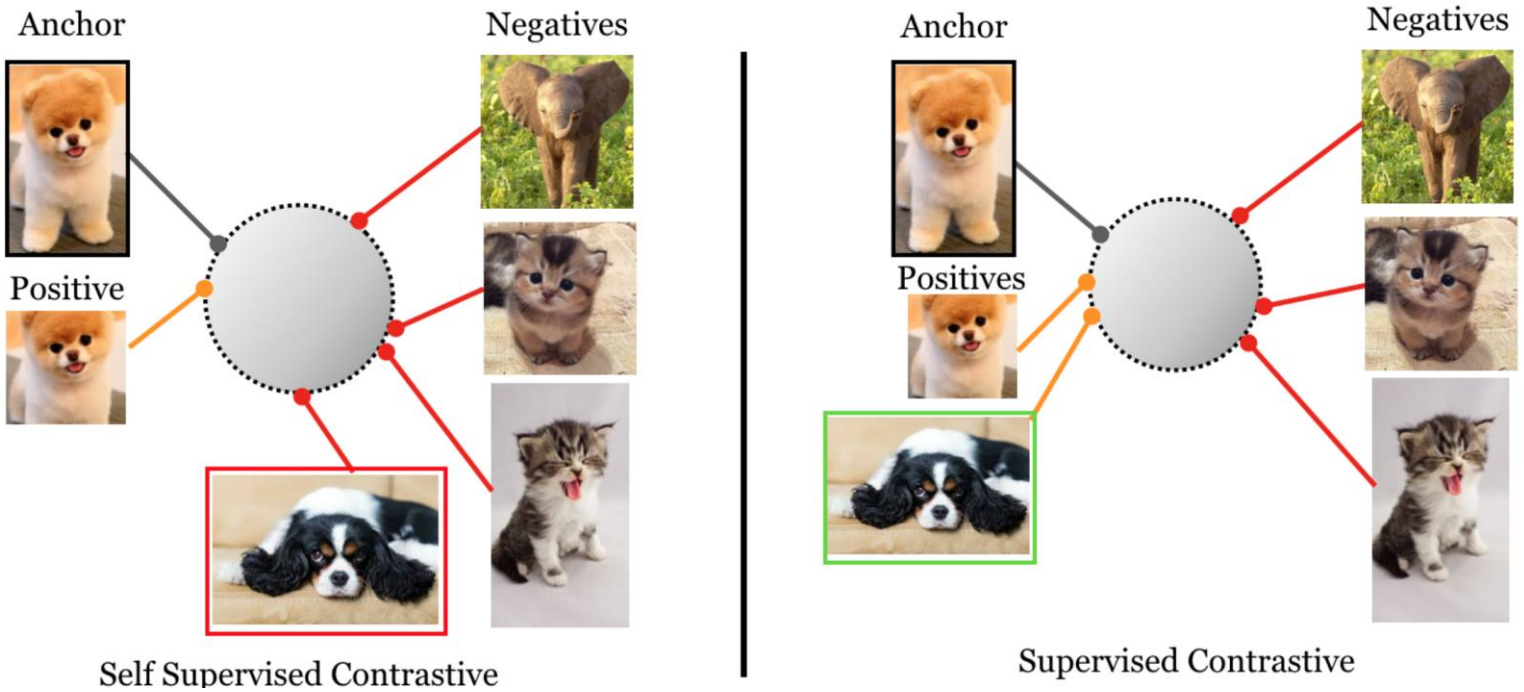



**Contrastive loss**
**Instances pairs**

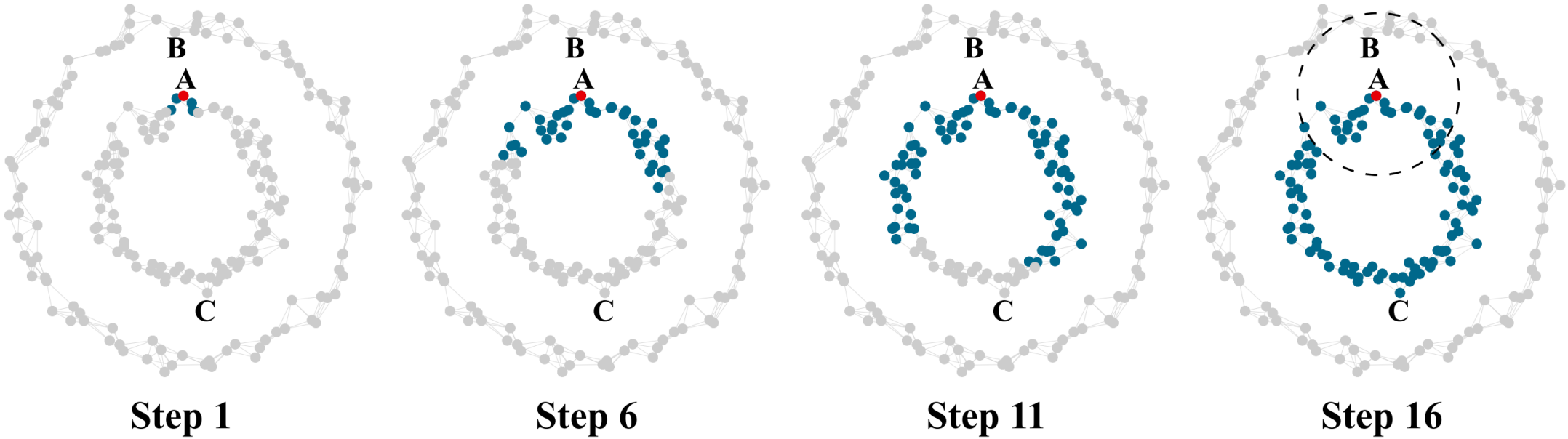
# Motivation

- **Instance-level**
- Contrastive learning aim to learn representations invariant to instance-level variations, which are provided by different views of the same instance.

- **Category-level**
- Learning representations invariant to category-level variations, which are provided by different instances from the same category.



# Positive Sample Discovery




**smoothness assumption**

if two points in a high-density region are close, then their semantic information should be similar.

# Positive Sample Discovery

The probability of sample $v_i$ being recognized as the j-th sample as:

$$P_{v_i}(j) = \frac{\exp(\bar{v}_j \cdot v_i/\tau)}{\sum_{k=1}^{n} \exp(\bar{v}_k \cdot v_i/\tau)}$$

The probability of $v_i$ being recognized as an image in S as:

$$P_{v_i}(\mathcal{S}) = \sum_{j \in \mathcal{S}} P_{v_i}(j)$$

The positive sample set N(i) of image $x_i$ as

$$\mathcal{N}(i) = \mathcal{N}_k(i) \cup \mathcal{N}_k(\mathcal{N}_k(i)) \cup \ldots \cup \underbrace{\mathcal{N}_k(\mathcal{N}_k(\mathcal{N}_k(\ldots \mathcal{N}_k}_{l}(i))))$$

*k is* small

# Hard Sampling Strategy

However, if we simply optimize the loss, the penalty strength on $Pv_i$ (j) for all $j \in N(i)$ is equal, which tends to maximize some easy optimized similarities.

Select P samples with the lowest similarity to construct the hard positive sample set $N_h(i)$.

## Hard negative samples set and Background samples set.

$$\mathcal{N}_{neg}(i) = \mathcal{N}_M(i) - \mathcal{N}(i)$$

$$B(i) = N_{neg}(i) \cup \mathcal{N}^h(i)$$

## Overall Loss Function

$$\mathcal{L}_{inv}(x_i) = -\log P_{v_i}(\mathcal{N}^h(i)|B(i))$$

$$= -\log \frac{\sum_{p \in \mathcal{N}^h(i)} \exp(\bar{v}_p \cdot v_i/\tau)}{\sum_{n \in B(i)} \exp(\bar{v}_n \cdot v_i/\tau)}$$

$$\mathcal{L}(x_i) = \mathcal{L}_{ins}(x_i) + \lambda_{inv} \cdot \omega(t) \cdot \mathcal{L}_{inv}(x_i)$$

Time function

$$\mathcal{L}_{ins}(x_i) = -\log P_{v_i}(i|\mathcal{N}_M(i) \cup \{i\})$$

# Experiments

Table 1: Linear classification results on ImageNet, Places205 and Pascal VOC07. We report 1-crop, top-1 accuracy. For ImageNet and VOC, we report the linear results for the output of the 16-th block. For Places205, we report the linear results for the output of the 15-th block. The BowNet has much more parameters than ordinary ResNet50 network because of the large size of fully connected layer.

| Method | Architecture | #Para | Epochs | ImageNet | Places | VOC |
|---|---|---|---|---|---|---|
| Supervised [28] | ResNet-50 | 26 | 200 | 75.9 | 51.5 | 87.5 |
| ***Self-supervised learning methods*** | | | | | | |
| Colorization [42] | ResNet-50 | 24 | 28 | 39.6 | 37.5 | 55.6 |
| Jigsaw [14] | ResNet-50 | 24 | 90 | 45.7 | 41.2 | 64.5 |
| Rotation [12] | ResNet-50 | 24 | 35 | 48.9 | 41.5 | 63.9 |
| BigBiGAN [9] | ResNet-50 | 24 | 488 | 56.6 | 49.8 | - |
| BoWNet(conv5) [11] | ResNet-50 | 65 | 280 | 60.5 | 50.1 | 78.4 |
| BoWNet(conv4) [11] | ResNet-50 | 65 | 280 | 62.1 | 51.1 | 79.3 |
| ***Methods based on contrastive learning*** | | | | | | |
| InsDis [39] | ResNet-50 | 24 | 200 | 54.0 | 45.5 | - |
| LocalAgg [46] | ResNet-50 | 24 | 200 | 58.8 | 49.1 | - |
| MoCo [16] | ResNet-50 | 24 | 200 | 60.6 | - | - |
| PIRL [28] | ResNet-50 | 24 | 800 | 63.6 | 49.8 | 81.1 |
| CMC [38] | ResNet-50-Lab | 47 | 400 | 64.1 | - | - |
| CPC [33] | ResNet-101 | 28 | - | 48.7 | - | - |
| CPC v2 [19] | ResNet-170 | 303 | - | 65.9 | - | - |
| AMDIM [1] | AMDIM | 626 | 150 | 68.1 | 55.0 | - |
| SimCLR [5] | ResNet-50-MLP | 28 | 1000 | 69.3 | - | 80.5 |
| MoCo v2 [6] | ResNet-50-MLP | 28 | 800 | 71.1 | - | - |
| PCL [26] | ResNet-50 | 24 | 200 | 62.2 | 49.2 | 82.2 |
| PCL [26] | ResNet-50-MLP | 28 | 200 | 65.9 | 49.8 | 84.0 |
| InvP (Ours) | ResNet-50 | 24 | 800 | 67.7 | 52.6 | 84.2 |
| InvP (Ours) | ResNet-50-MLP | 28 | 800 | **71.3** | **53.5** | **84.7** |

# Experiments

Table 2: Semi-supervised learning performance on ImageNet. We fine-tune our pre-trained models on 1% or 10% of ImageNet labeled data sampled from training set. We report top-5 accuracy on the held-out validation set. The results of other methods are adopted from original papers.

| Method | Architecture | Pretrain Epochs | Top5 Accuracy 1% | 10% |
|---|---|---|---|---|
| ***Semi-supervised learning methods*** | | | | |
| VAT + Ent Min [15, 29] | ResNet-50v2 | - | 47.0 | 83.4 |
| $S^4L$ Exemplar [41] | ResNet-50v2 | - | 47.0 | 83.7 |
| $S^4L$ Rotation [41] | ResNet-50v2 | - | 53.4 | 83.8 |
| LLP [45] | ResNet-50 | - | 61.9 | 88.5 |
| ***Unsupervised learning methods*** | | | | |
| Jigsaw [14] | ResNet-50 | 90 | 45.3 | 79.3 |
| InsDis [39] | ResNet-50 | 200 | 39.2 | 77.4 |
| PIRL [28] | ResNet-50 | 800 | 57.2 | 83.8 |
| SimCLR [5] | ResNet-50-MLP | 1000 | 75.5 | 87.8 |
| PCL [26] | ResNet-50 | 200 | 75.6 | 86.2 |
| InvP (Ours) | ResNet-50 | 800 | 76.7 | 87.2 |
| InvP (Ours) | ResNet-50-MLP | 800 | **78.2** | **88.7** |

# Experiments

Table 3: Transfer learning performance on different datasets. We compare our method with SimCLR [5], supervised model and model trained from scratch. We report top-1 accuracy for CIFAR10, CIFAR100 and Stanford Cars; mean per-class accuracy for Caltech-101, Oxford-IIIT Pets and Oxford 102 Flowers; and the 11 point mAP for Pascal VOC2007, which is same as the setting of SimCLR

| **Method** | CIFAR10 | CIFAR100 | VOC | Caltech101 | Cars | Pets | Flowers |
|---|---|---|---|---|---|---|---|
| Scratch | 95.9 | 80.2 | 67.3 | 72.6 | 91.4 | 81.5 | 92.0 |
| Supervised | 97.5 | 86.4 | 85.0 | 93.3 | 92.1 | 92.1 | 97.6 |
| SimCLR | 97.7 | **85.9** | 84.1 | 92.1 | **91.3** | 89.2 | **97.0** |
| InvP (Ours) | **97.9** | 84.7 | **85.4** | **92.5** | 90.3 | **89.4** | 96.7 |

# Experiments

Table 4: The results of object detection. We fine-tune the unsupervised model on the Pascal VOC2007+2012 training set and report $AP_{50}$, $AP_{75}$ and $AP_{all}$ on VOC2007 test set, which is a widely adopted setting [16, 28, 14]. The proposed method outperforms other competitors.

| Method | Dataset | Network | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|---|---|
| Supervised | ImageNet-1k | R50 C4 | 53.2 | 80.8 | 58.5 |
| Jigsaw [14, 28] | ImageNet-22k | R50-C4 | 48.9 | 75.1 | 52.9 |
| InsDis [39] | ImageNet-1k | R50-C4 | 52.3 | 79.1 | 56.9 |
| MoCo [16] | ImageNet-1k | R50-C4 | 55.2 | 81.4 | 61.2 |
| MoCo [16] | ImageNet-1k | R50-C5 | 53.8 | 81.1 | 58.6 |
| PIRL [28] | ImageNet-1k | R50-C4 | 54.0 | 80.7 | 59.7 |
| BoWNet [11] | ImageNet-1k | R50-C4 | 55.8 | 81.3 | 61.1 |
| MoCo v2 [6] | ImageNet-1k | R50-C4 | **57.4** | **82.5** | **64.0** |
| InvP (Ours) | ImageNet-1k | R50-C4 | 56.2 | 81.8 | 61.5 |

Table 5: Results of ablation study on ImageNet linear classification. We train all models with 200 epochs and report the top-1 center-crop accuracy. The backbone is ResNet-50 without MLP head.

| | InvP | KNN | Without Hard Positive | Without Hard Negative |
|---|---|---|---|---|
| Acc | 63.3 | 57.6 | 60.7 | 61.9 |

# Experiments

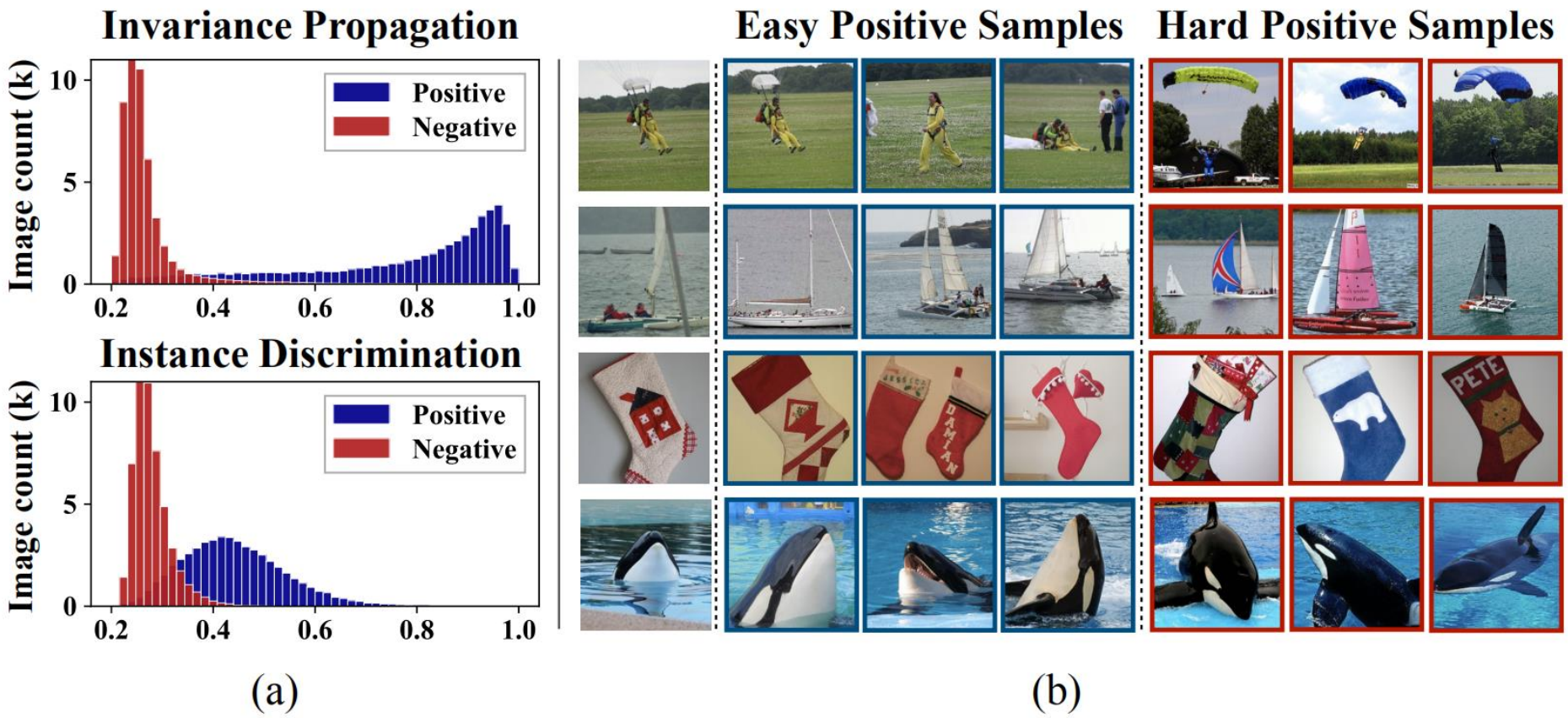



Figure 2: (a): The distribution of positive and negative similarities. We compare our model with the instance discrimination model [39]. (b): Visualization of the positive samples. The first column represents the anchor samples (query). For each anchor sample, we give comparison between the easy positive samples and the hard positive samples.

# Thanks