WHAT MAKES INSTANCE DISCRIMINATION GOOD FOR TRANSFER LEARNING?

Anonymous authors
Paper under double-blind review

A1 EFFECTS OF PRETRAINING AND FINETUNING ITERATIONS

We also conduct experiments to examine the effects of pretraining optimization epochs and finetuning iterations. We show results in Figure 1, and find that longer optimization during pretraining consistently improves detection transfer for both supervised and unsupervised models. This suggests that overfitting is not an issue for either pretraining method. Unsupervised pretraining is seen to converge much faster during pretraining, and supervised pretrained models tend to converge faster in the initial iterations of detection finetuning but may not converge optimally.

We notice that supervised pretraining benefits from more optimization epochs. To explore the limit of supervised pretraining, we investigate larger numbers of supervised pretraining epochs. In Table 1, supervised pretraining continues to improve performance until 800 epochs, but may suffer from overfitting as indicated by the performance on ImageNet classification. For detection transfer, the improved supervised pretraining still falls short MoCo on AP and AP_{75}, while it outperforms MoCo on AP_{50}. This may possibly be due to the superior semantic classification ability of supervised models. Further discussion of the results are beyond the scope of the paper.

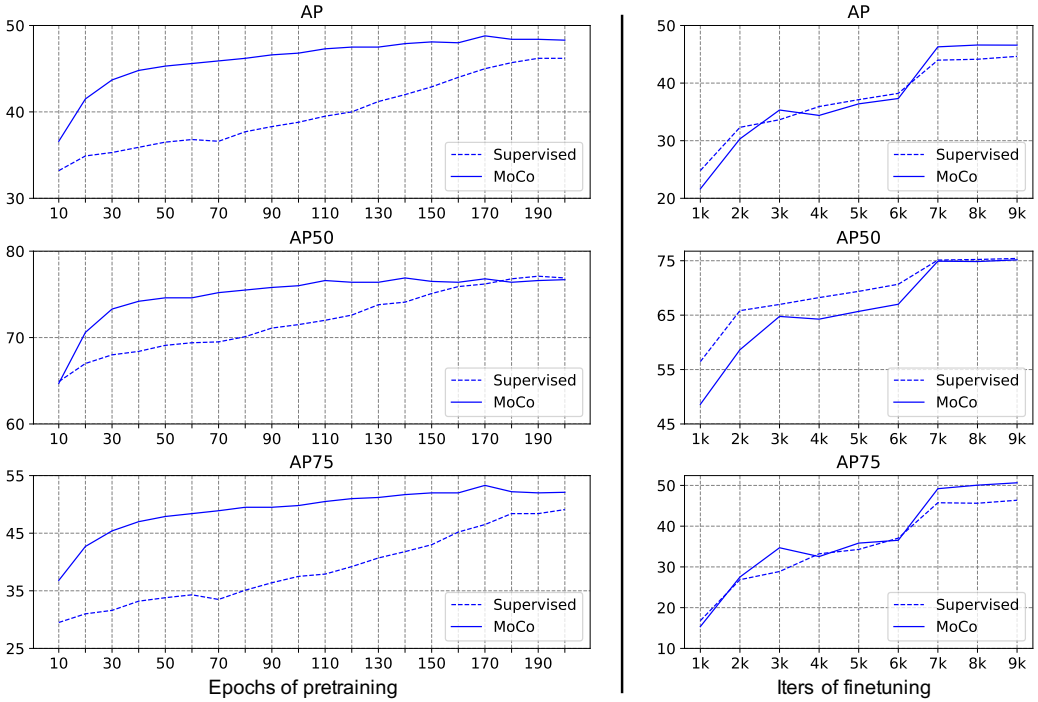

Figure 1: Performance at intermediate pretraining checkpoints and finetuning checkpoints.

A2 EFFECTS OF IMAGE AUGMENTATIONS ON PRETRAINING

We show full results of object detection on PASCAL VOC07, object detection and instance segmentation on MSCOCO, and semantic segmentation on Cityscapes in Table 2.

Table 1: Longer supervised pretraining for object detection transfer on PASCAL VOC.

Pretraining Epochs	ImageNet Acc	VOC07 detection AP	VOC07 detection AP_{50}	VOC07 detection AP_{75}	VOC0712 detection AP	VOC0712 detection AP_{50}	VOC0712 detection AP_{75}
90	75.5	45.4	76.3	47.0	54.8	82.1	60.4
200	77.3	46.0	76.7	48.3	55.4	82.3	61.6
400	77.8	47.7	78.0	50.7	56.1	82.9	62.8
800	77.7	47.6	77.5	51.0	56.4	82.7	62.9
MoCo	67.5	48.5	76.8	52.7	56.9	82.2	63.5

Table 2: The effects of pretraining image augmentations on the transfer performance for supervised and unsupervised models.

Pytorch Augmentation	Supervised				Unsupervised			
	ImageNet	VOC07 detection			ImageNet	VOC07 detection		
	Acc	AP	AP_{50}	AP_{75}	Acc	AP	AP_{50}	AP_{75}
+ RandomHorizontalFlip(0.5)	70.9	43.4	74.0	44.5	6.4	32.3	58.3	31.4
+ RandomResizedCrop(224)	77.5	45.5	76.2	47.4	53.0	43.2	71.2	45.4
+ ColorJitter(0.4, 0.4, 0.4, 0.1)	77.4	45.9	76.7	48.0	62.7	45.7	74.4	48.6
+ RandomGrayscale(p=0.2)	77.7	46.4	77.3	49.0	66.4	47.7	76.0	51.5
+ GaussianBlur(0.1, 0.2)	77.3	46.2	76.8	48.9	67.5	48.5	76.8	52.7

Supervised						Unsupervised					
COCO detection			COCO segmentation			COCO detection			COCO segmentation		
AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
38.6	58.5	41.7	33.7	55.1	35.9	34.2	52.7	36.7	30.6	49.9	32.4
38.9	59.3	41.6	34.0	55.7	36.0	36.8	56.1	39.7	32.3	52.9	34.4
39.3	59.6	42.3	34.4	56.1	36.4	37.5	56.9	40.5	33.0	54.0	35.0
39.1	59.2	42.0	34.2	55.6	36.4	38.6	58.0	41.9	33.8	54.8	36.0
38.9	59.1	41.8	33.9	55.4	35.9	38.7	58.1	42.0	34.0	55.1	36.4

Supervised			Unsupervised		
Cityscapes Segmentation					
mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
78.0	85.2	96.0	72.7	81.3	95.3
78.7	85.6	96.1	76.6	84.2	95.9
78.7	85.9	96.1	77.7	85.2	96.0
78.7	85.6	96.1	78.4	85.7	96.1
78.8	85.8	96.1	78.6	85.7	96.2

A3 Effects of Dataset Semantics on Pretraining

We report full transfer performance with pretraining on various datasets in Table 3. We also provide a visualization of various datasets for training these models in Figure 2.

A4 Details on Image Reconstruction by Inverting Features

A4.1 Method Details

We use the same architecture for the reconstruction network $r_\theta(\cdot)$ as in the original deep image prior paper. It is an encoder-decoder network with the following architecture. Let C_k^m denote a Convolution-BatchNorm-LeakyReLU layer with k channels and $m \times m$ spatial filters; CD_k^m denote a Convolution-Downsample-BatchNorm-LeakyReLU layer, and CU_k^m denote a Convolution-BatchNorm-LeakyReLU-Upsample layer. We use a stride of 2 for both the upsampling and downsampling layers.

Table 3: Transfer performance with pretraining on various datasets. "ImageNet-10%" denotes subsampling 1/10 of the images per class on the original ImageNet. "ImageNet-100" denotes subsampling 100 classes in the original ImageNet. Supervised pretraining uses the labels in the corresponding dataset, and unsupervised pretraining follows MoCo-v2. Supervised models for CelebA and Places are trained with identity and scene categorization supervision, while supervised models for COCO and Synthia are trained with semantic bounding box and segmentation supervision for detection and segmentation networks, respectively.

Pretraining Data	#Imgs	Annotation	Supervised				Unsupervised			
			ImageNet	VOC07 detection			ImageNet	VOC07 detection		
			Acc	AP	AP_{50}	AP_{75}	Acc	AP	AP_{50}	AP_{75}
ImageNet	1281K	object	77.3	46.2	76.8	48.9	67.5	48.5	76.8	52.7
ImageNet-10%	128K	object	57.8	42.4	73.5	43.1	58.9	45.5	74.4	48.0
ImageNet-100	124K	object	50.9	42.0	72.4	43.3	56.5	45.6	73.9	48.5
Places	2449K	scene	52.3	39.1	70.0	38.7	57.1	46.7	74.9	50.2
CelebA	163K	identity	30.3	37.5	66.1	36.9	40.1	45.3	72.4	48.4
COCO	118K	bbox	57.8	53.3	80.3	59.5	50.6	46.1	74.5	49.4
Synthia	365K	segment	30.2	40.2	70.3	40.2	13.5	37.4	65.0	37.2

Supervised						Unsupervised					
COCO detection			COCO segmentation			COCO detection			COCO segmentation		
AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
38.9	59.1	41.8	33.9	55.4	35.9	38.7	58.1	42.0	34.0	55.1	36.4
37.7	57.5	40.5	33.1	54.3	35.1	38.6	58.0	41.7	33.9	54.9	36.0
37.1	56.6	40.1	32.5	53.3	34.5	38.3	57.7	41.6	33.6	54.5	35.5
36.6	56.3	39.1	32.2	53.1	34.1	38.4	58.0	41.3	33.6	54.5	35.7
36.4	55.5	39.4	32.2	52.2	34.5	37.5	56.5	40.3	33.0	53.5	35.3
39.1	58.9	42.3	34.0	55.5	36.2	38.4	58.0	41.6	33.7	54.6	35.8
37.3	57.1	40.4	32.9	53.8	35.0	36.1	55.0	38.6	31.7	51.9	33.7

Supervised			Unsupervised		
Cityscapes Segmentation					
mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
78.8	85.8	96.1	78.6	85.7	96.2
77.7	85.0	96.0	78.1	85.6	96.1
77.0	84.5	95.9	77.8	85.2	96.1
77.6	85.0	96.0	78.8	86.2	96.1
76.5	84.3	95.9	76.8	84.4	95.9
78.3	85.5	96.0	78.3	85.6	96.1
76.5	84.1	95.9	75.6	83.6	95.8

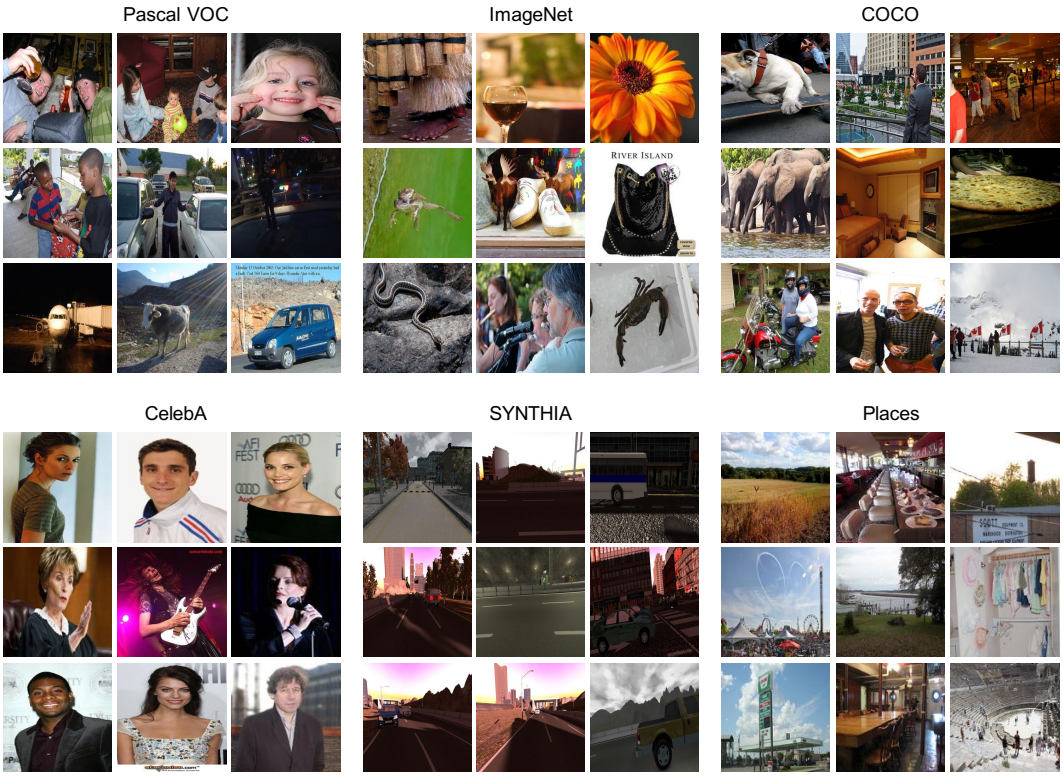

Figure 2: Example images of various datasets used for the pretraining study.

Encoder: $CD^7_{16} - C^7_{16} - CD^7_{32} - C^7_{32} - CD^5_{64} - C^5_{64} - CD^5_{128} - C^5_{128} - CD^3_{128} - C^3_{128} - CD^3_{128} - C^3_{128}$

Decoder: $C^7_{16} - CU^7_{16} - C^7_{32} - CU^7_{32} - C^5_{64} - CU^5_{64} - C^5_{128} - CU^5_{128} - C^3_{128} - CU^3_{128} - C^3_{128} - CU^3_{128}$

The input $z_0 \in R^{H \times W \times 32}$ is initialized with uniform noise between 0 and 0.1. For each image, the optimization takes 3000 iterations of an Adam optimizer with a learning rate of 0.001.

A4.2 Evaluating Reconstructions by Perceptual Metrics

To measure the reconstruction quality quantitatively, we calculate the perceptual distance between the reconstruction and the input image, using a deep learning based approach (Zhang et al. (2018)) with a SqueezeNet network. We randomly select one image per class from the ImageNet validation set for 1000 images in total. The average distance of reconstructions using MoCo is 5.59, while it is 6.43 for the supervised network. We provide a scatter plot of perceptual distance from individual reconstructions. In Figure 3, we can see that the reconstructions generated by MoCo are generally closer to the original images than those generated by the supervised method.

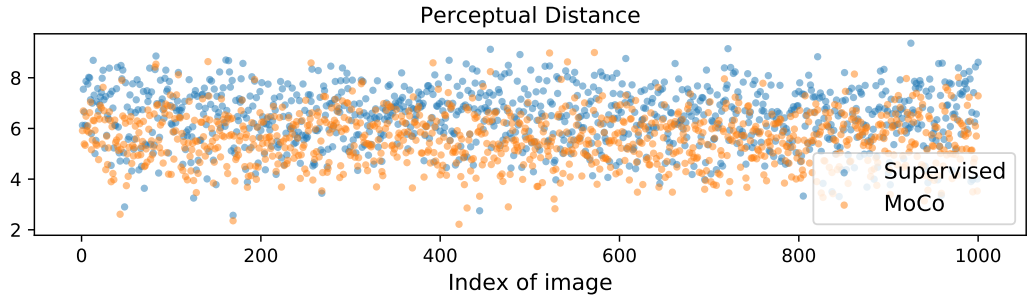

Figure 3: Perceptual distance between the reconstruction and the original image on 1000 validation images.

A5 MORE RESULTS ON EXEMPLAR-BASED SUPERVISED PRETRAINING

We show full transfer performance of our proposed Exemplar-based supervised pretraining in Table 4.

Table 4: Exemplar-based supervised pretraining which does not enforce explicit constraints on the positives. It shows consistent improvements over the MoCo baselines by using labels.

Methods	ImageNet	VOC07 detection			Cityscapes segmentation		
	Acc	AP	AP_{50}	AP_{75}	mIoU	mAcc	aAcc
MoCo-v1	60.8	46.6	74.9	50.1	78.4	85.6	96.1
Exemplar-v1	64.6	47.2	76.0	50.6	78.9	86.0	96.2
MoCo-v2	67.5	48.5	76.8	52.7	78.6	85.7	96.2
Exemplar-v2	68.9	48.8	77.2	53.1	78.8	85.9	96.2

Methods	COCO detection			COCO segmentation		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
MoCo-v1	38.5	58.3	41.6	33.6	54.8	35.6
Exemplar-v1	39.0	58.7	42.0	34.0	55.4	36.3
MoCo-v2	38.7	58.1	42.0	34.0	55.1	36.4
Exemplar-v2	39.4	59.1	42.7	34.4	55.9	36.5

Since our Exemplar pretraining uses a different set of parameters from MoCo, we provide an ablation study over the parameter k and τ for ImageNet linear readout in Table 5.

Table 5: An ablation study of parameter k and τ for MoCo and Exemplar pretraining.

Methods	k	τ	ImageNet acc
MoCo-v1	65536	0.07	60.8
MoCo-v1	1M	0.07	60.9
Exemplar-v1	1M	0.07	64.6
Exemplar-v1	1M	0.1	63.9
MoCo-v2	65536	0.2	67.5
MoCo-v2	1M	0.1	66.9
MoCo-v2	1M	0.2	67.8
Exemplar-v2	1M	0.07	68.1
Exemplar-v2	1M	0.1	68.9
Exemplar-v2	1M	0.2	67.9

A6 ADDITIONAL RESULTS OF DIAGNOSING DETECTION ERROR

We provide a full analysis over 20 object categories on the VOC07 test set. For each category, a pie chart is given to show the distribution of four kinds of errors in top-ranked false positives. For each category, the false positives are chosen to be within the top N detections, where N is chosen to be the number of ground truth objects in each category. The four types of false positives include: poor localization (Loc), confusion with similar objects (Sim), confusion with other VOC objects (Oth), or confusion with background or unlabeled objects (BG). In Figure 4, we compare the error distribution between the MoCo results and supervised results. It is apparent that detection results from the MoCo pretrained model exhibits a smaller proportion of localization errors.

REFERENCES

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

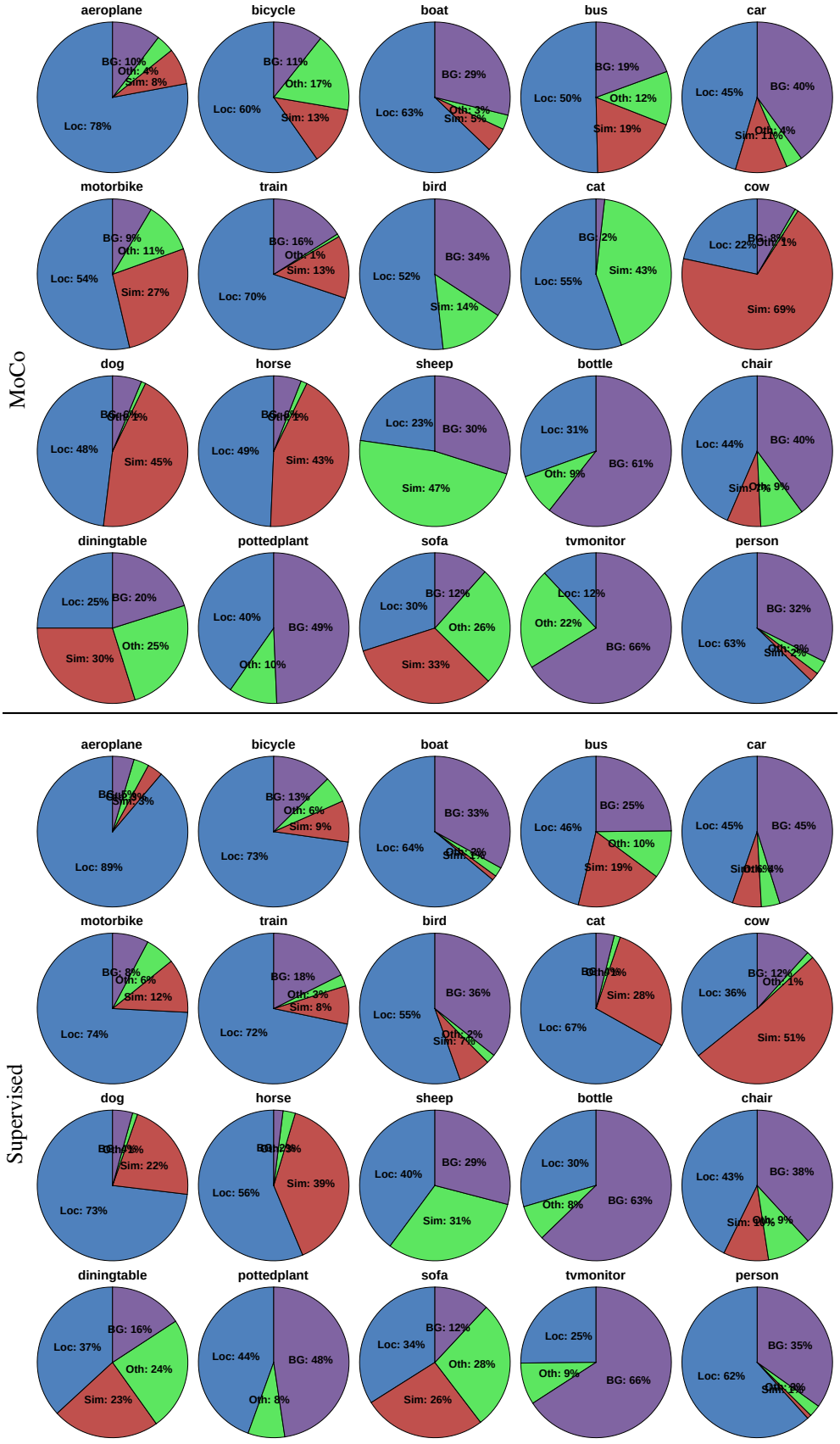

Figure 4: Distribution of four types of false positives for each category.