What makes instance discrimination good for transfer learning?

Nanxuan Zhao\(^1\star\)  Zhirong Wu\(^2\star\)  Rynson W.H. Lau\(^1\)  Stephen Lin\(^2\)

\(^1\)City University of Hong Kong \(^2\)Microsoft Research Asia

Abstract

Unsupervised visual pretraining based on the instance discrimination pretext task has shown significant progress. Notably, in the recent work of MoCo, unsupervised pretraining has shown to surpass the supervised counterpart for finetuning downstream applications such as object detection on PASCAL VOC. It comes as a surprise that image annotations would be better left unused for transfer learning. In this work, we investigate the following problems: What makes instance discrimination pretraining good for transfer learning? What knowledge is actually learned and transferred from unsupervised pretraining? From this understanding of unsupervised pretraining, can we make supervised pretraining great again? Our findings are threefold. First, what truly matters for this detection transfer is low-level and mid-level representations, not high-level representations. Second, the intra-category invariance enforced by the traditional supervised model weakens transferability by increasing task misalignment. Finally, supervised pretraining can be strengthened by following an exemplar-based approach without explicit constraints among the instances within the same category.

1 Introduction

Recently, a remarkable transfer learning result with unsupervised pretraining was reported on visual object detection. The pretraining method MoCo [12, 5] established a milestone by outperforming the supervised counterpart, with an AP of 46.6 compared to 42.4 on VOC07. Supervised pretraining has been the de facto standard for finetuning downstream applications, and it is surprising that labels of one million images, which took years to collect, appear to be unhelpful and perhaps even harmful for transfer learning. This raises the question of why unsupervised pretraining provides better transfer performance and supervised pretraining fails short.

The leading unsupervised pretraining methods follow an instance discrimination pretext task [7, 35, 12, 40, 3], where the features of each instance are pulled away from those of all other instances in the training set. Invariances are encoded from low-level image transformations such as cropping, scaling and color jittering. With such low-level induced invariances [12, 4], strong generalization has been achieved to high-level visual concepts such as object categories on ImageNet. On the other hand, the widely adopted supervised pretraining method optimizes the cross-entropy loss over the predictions and the labels. As a result, training instances within the same category are drawn closer while the training instances of different categories are pulled apart.

Toward a deeper understanding of why unsupervised pretraining by instance discrimination performs so well, we dissect the performance of both unsupervised and supervised pretraining methods on

\(^1\)In this paper, the term unsupervised pretraining specifically refers to MoCo [12, 5], as it is the state-of-the-art and a good representative for unsupervised learning. Other unsupervised learning approaches such as GANs and autoencoders are not considered in this work.

\(^\star\)contributed equally.

Preprint. Under review.
the downstream task of object detection. Our study begins by examining the common belief that
the transfer of high-level semantic information is the key to effective transfer learning [11, 20]. On
unsupervised pretraining with different types of image sets, it is found that transfer performance
is largely unaffected by the high-level semantic content of the pretraining data, whether it matches
the semantics of the target data or not. Moreover, pretraining on synthetic data, whose low-level
properties are inconsistent with real images, leads to a drop in transfer performance. These results
indicate that it is primarily low-level and mid-level representations that are transferred.

We also investigate how well supervised and unsupervised pretraining are aligned to the object
detection task. First, the detection errors of both are diagnosed using the detection toolbox [14]. It is
found that supervised pretraining is more susceptible than unsupervised pretraining to localization
error. Secondly, to understand the localization error, we examine how effectively images can be
reconstructed from supervised and unsupervised representations. The results show that supervised
representations mainly model the discriminative parts of objects, in contrast to the more holistic
modeling of unsupervised representations pretrained to discriminate instances rather than classes.
Both sets of experiments suggest that there exists a greater misalignment of supervised pretraining to
the downstream task of object detection, which requires accurate localization and full delineation of
the object region.

Based on these studies, we conclude that, in visual object detection, not only it is unnecessary to
transfer high-level semantic information, but learning to discriminate among classes is misaligned
with object detection. We thus hypothesize that the essential difference that makes supervised
pretraining weaker (and instance discrimination stronger) is the common practice of minimizing
intra-class variation. The crude assumption that all instances within one category should be alike in
the feature space neglects the unique information from each instance that may have significance in
downstream applications. This motivates us to explore a new supervised pretraining method that does
not explicitly embed instances of the same class in close proximity of one another. Rather, we follow
exemplar SVM [21] in pulling away the true negatives of each training instance without enforcing
any constraint on the positives. This respects the data distribution in a manner that preserves the
variations in the positives, and our new pretraining method is shown to yield consistent improvements
for both ImageNet classification and object detection transfer.

We expect these findings to have broad implications over a variety of transfer learning applications.
As long as there exists any misalignment between the pretraining and downstream tasks (which is
ture of most transfer learning scenarios in computer vision), one should always be careful about
overfitting to the supervised invariances defined by the pretraining labels. We further test on two
other transfer learning scenarios: few-shot image recognition and facial landmark prediction. Both of
them are found to align with the conclusions obtained from our object detection study.

## 2 A Case Study on Object Detection

We study the transfer performance of a variety of pretrained models for object detection on PASCAL
VOC07. Given a pretrained network, following the MoCo paper, we finetune all layers in the network
with synchronized batch normalization. Optimization takes 9k iterations on 8 GPUs with a batch
size of 2 images per GPU. The learning rate is initialized to 0.02 and decayed to be 10 times smaller
after 6k and 8k iterations. We use the detectron2 codebase [34] and the ResNet50-C4 architecture
as the backbone in the Faster R-CNN framework. Detection training is performed on the VOC07
train/val combined splits, and the performance is evaluated on the VOC07 test set using COCO-style
AP metrics. All results are reported as the average of three independent runs.

### 2.1 Comparison of Detection Transfer

We begin by confirming the advantage of unsupervised pretraining by conducting detailed compar-
isons to supervised pretraining on common ground. In setting the common ground, we account for
pretraining image augmentations, pretraining optimization epochs and finetuning iterations. This
comparison is designed to check whether supervised pretraining could have performed better if the
model were more overfitted or the model were pretrained with a different set of image augmentations.

We use MoCo-v2 [5] for unsupervised pretraining and the standard cross entropy loss for supervised
pretraining with ResNet50. Both types of pretraining are optimized for 200 epochs with cosine

2
Table 1: Detection comparisons of supervised and unsupervised pretraining on common ground.

| Pytorch Augmentation          | Supervised | | | | | | Unsupervised | | | | |
|------------------------------|------------|---|---|---|---|---|---|---|---|---|---|---|---|
|                              | ImageNet   | VOC07 detection | Acc | AP | AP$_{50}$ | AP$_{75}$ | ImageNet | VOC07 detection | 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 |

(a) Image augmentations for pretraining.

(b) Performance at intermediate pretraining checkpoints and finetuning checkpoints.

First, in Table 1 (a), supervised pretraining is found to also benefit from image augmentations such as color jittering and random grayscaling, but it is negatively affected by Gaussian blurring to a small degree. However, the improved supervised pretraining model still falls short of the unsupervised pretrained model (with almost equal performance at AP$_{50}$). Second, in Table 1 (b), 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. Results of even longer supervised pretraining are included in the supplement [A1].

These results confirm that unsupervised pretraining outperforms supervised pretraining for detection transfer on common ground. We analyze why this happens in the following sections.

2.2 Effect of Dataset Semantics on Pretraining

The strong performance of unsupervised pretraining on the linear classification protocol for ImageNet [12, 8] shows that the features capture high-level semantic representations of object categories. In supervised pretraining, it is a common belief that this high-level representation [11, 20] is what transfers from ImageNet to VOC detection. Here, we challenge this conclusion by studying the

---

2We optimally set the scale parameter to 0.08 for supervised pretraining and 0.2 for unsupervised pretraining.
Table 2: 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  | ImageNet | VOC07 |
|                  |       |            | Acc       | AP    | AP     | AP    | AP    |
|                  |       |            | 50        | 50    | 75     | 75    |
| ImageNet         | 1281K | object     | 77.3      | 46.2  | 76.8   | 48.9  |
| ImageNet-10%     | 128K  | object     | 57.8      | 42.4  | 73.5   | 43.1  |
| ImageNet-100     | 124K  | object     | 50.9      | 42.0  | 72.4   | 43.3  |
| Places           | 2449K | scene      | 52.3      | 39.1  | 70.0   | 38.7  |
| CelebA           | 163K  | identity   | 30.3      | 37.5  | 66.1   | 36.9  |
| COCO             | 118K  | bbox       | 57.8      | 53.3  | 80.3   | 59.5  |
| Synthia          | 365K  | segment    | 30.2      | 40.2  | 70.3   | 40.2  |

Transfer of unsupervised models pretrained on images with little or no semantic overlap with the VOC dataset. These image datasets include faces, scenes, and synthetic street-view images. We also investigate how the size of the unsupervised pretraining dataset affects transfer performance.

We follow the MoCo-v2 pretraining method and optimize the unsupervised network for 200 epochs. For the smaller datasets, we increase the number of training epochs and maintain the effective number of optimization iterations. The results are summarized in Table 2. It can be seen that the transfer learning performance of unsupervised pretraining is relatively unaffected by the pretraining image data, while supervised pretraining depends on the supervised semantics. All the supervised networks are negatively impacted by the change of pretraining data except when the annotation contains pixel-level supervision as in COCO for bounding box detection and Synthia for semantic segmentation. Also, with the smaller amounts of pretraining data in ImageNet-10% and ImageNet-100, the advantage of unsupervised pretraining becomes more pronounced in relation to supervised models, which suggests stronger ability for generalization with less data. Results on VOC0712 are presented in the supplement A2.

Unsupervised pretraining on faces and scenes achieves almost the same transfer results as pretraining on ImageNet. Since the face dataset has almost no semantic overlap with the VOC objects (besides the human category), transfer of high-level representations can be seen as extraneous. We further test unsupervised pretraining on the synthetic dataset Synthia [23], which exhibits low-level statistics different from real images. With this model, there is a substantial performance drop. We can therefore conclude that instance discrimination pretraining mainly transfers low-level and mid-level representations. In Table 2, we also test the linear readoff on ImageNet-1K classification using various pretrained models. These models perform very differently, suggesting that the last-layer features learned by unsupervised training still overfit to the training data semantics.

2.3 Task Misalignment and Information Loss

A strong high-level representation is not necessary for effective transfer to object detection, but this itself does not explain why unsupervised pretraining yields better performance than supervised pretraining. We notice that a larger performance gap exists on AP_{75} than on AP_{50}, which suggests that supervised pretraining is weaker at precise localization. For additional analysis, we use the detection toolbox [14] to diagnose detection errors. Figure 1 compares the error distributions of the supervised and unsupervised transfer results on three example categories. We find that the detection errors of supervised pretraining models are more frequently the result of poor localization, where low IoU bounding boxes are mistakenly taken as true positives. A comprehensive comparison for all categories is provided in the supplement A5.

For further examination, we compare image reconstruction from the features of supervised and unsupervised pretraining. This reconstruction is performed by inverting the layer4 features (dimension...
Figure 1: Analyzing detection error using the detection toolbox [14]. Distribution of top-ranked false positive (FP) types for finetuning with supervised and unsupervised methods. Supervised pretraining models more frequently result in localization errors than unsupervised pretraining models. Each FP is categorized into 1 of 4 types: Loc—poor localization; Sim—confusion with a similar category; Oth—confusion with a dissimilar object category; BG—a FP that fires on background.

Figure 2: Image reconstruction by feature inversion. We use the method of deep image prior [30] to reconstruct images by a pretrained network. Unsupervised features allow for holistic reconstruction over the entire image, while supervised features lose information in many regions.

of $7 \times 7 \times 2048$) using the deep image prior [30]. Specifically, given an image input $x_0$, we optimize a reconstruction network $r_\theta$ to produce a reconstruction $x$ that is close to the input $x_0$ in the embedding space of a pretrained encoding network $f(\cdot)$, 

$$\min_\theta E(f(x), f(x_0)), \quad x = r_\theta(z_0). \quad (1)$$

The input $z_0$ to the reconstruction network $r_\theta(\cdot)$ is fixed spatial noise, and the distance function $E(\cdot)$ is implemented as the $L2$ distance. The architecture of $r_\theta(\cdot)$ is an autoencoder network with six blocks for both the encoder and decoder, as detailed in the appendix. With this inversion method, we observe how well a pretrained network $g(\cdot)$ can recover image pixels from the features.

We visualize the reconstructions for both the supervised and unsupervised pretrained networks in Figure[2]. It is apparent that the unsupervised network provides more complete reconstructions, while the supervised network loses information over large regions in the images, likely because its features are mainly attuned to the most discriminative object parts, which are central to the classification task, rather than objects and images as a whole. The resulting loss of information may prevent the supervised network from detecting the full envelope of the object. To measure the reconstruction quality quantitatively, we calculate the perceptual distance between the reconstruction of each method and the input image, using a deep learning based approach [38] 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.

In Figure[2] we notice that from the unsupervised network features the images are reconstructed at the correct scale and location. Though instance discrimination encodes invariances through spatial and scale transformations, features learned this way are still sensitive to these factors [3]. A possible explanation is that in order to make one instance unique from all other instances, the network strives to preserve as much information as possible.
Table 3: 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 Acc | VOC07 detection | VOC0712 detection |
|---------------|--------------|-----------------|-------------------|
|               |              | AP   | AP50 | AP75 | AP   | AP50 | AP75 |
| MoCo-v1       | 60.8         | 46.6 | 74.9 | 50.1 | 55.9 | 81.5 | 62.6 |
| Exemplar-v1   | 64.6         | 47.2 | 76.0 | 50.6 | 56.3 | 81.9 | 62.8 |
| MoCo-v2       | 67.5         | 48.5 | 76.8 | 52.7 | 57.0 | 82.4 | 63.6 |
| Exemplar-v2   | 68.9         | 48.8 | 77.2 | 53.1 | 57.2 | 82.7 | 63.7 |

3 A Better Supervised Pretraining Method

Annotating one million images in ImageNet provides rich semantic information which could be useful for downstream applications. However, traditional supervised learning minimizes intra-class variation by optimizing the cross-entropy loss between predictions and labels. By doing so, it focuses on the discriminative regions [27] within a category but at the cost of information loss in other regions. A better supervised pretraining method should instead pull away features of the true negatives for each instance without enforcing explicit constraints on the positives. This preserves the unique information of each positive instance while utilizing the label information in a weak manner.

We propose a new supervised pretraining method inspired by exemplar SVM [21], which trains an individual SVM classifier to separate each instance from its negatives. Unlike the original exemplar SVM which represents positives non-parametrically and negatives parametrically, our pretraining scheme models all instances in an non-parametric fashion in a spirit similar to instance discrimination [35]. Concretely, we follow the framework of momentum contrast [12], where each training instance \(x_i\) is augmented twice to form \(x^q_i\) and \(x^k_i\), which are fed into two encoders for embedding, \(v_i = f_q(x^q_i), k_i = f_k(x^k_i)\). Please refer to MoCo [12] for details about the momentum encoders. But instead of discriminating from all other instances [35], the loss function uses the labels \(y_i\) to filter the true negatives,

\[
L_{q_i} = -\log \frac{\exp(q_i^T k_i / \tau)}{\sum_{y_j \neq y_i} \exp(q_i^T k_j / \tau)}, \tag{2}
\]

where \(\tau\) is the temperature parameter. We set the baselines to be MoCo-v1 and MoCo-v2, and denote our corresponding methods as Exemplar-v1 and Exemplar-v2. Temperature \(\tau = 0.07\) is used for Exemplar-v1, and \(\tau = 0.1\) is used for Exemplar-v2 with ablations in the supplement [4].

Experimental results are presented in Table 3. By filtering true negatives using semantic labels, our method consistently improves classification performance on ImageNet and transfer performance for object detection. This is in contrast to traditional supervised learning, where ImageNet performance is improved and its transfer performance is compromised. We note that the ImageNet classification performance of our exemplar-based training, 68.9%, is still far from the traditional supervised learning result of 77.3%. This leaves room for future research on even better classification and transfer learning performance.

4 Implications for Other Transfer Learning Scenarios

The presented studies focus on the transfer scenario of ImageNet pretraining to VOC object detection. For other downstream applications, the nature of the task misalignment may differ. Thus, we additionally consider two other transfer learning scenarios to study the implications of overfitting to the supervised pretraining semantics and how it can be improved by our exemplar-based pretraining.

4.1 Few-shot Recognition

The first scenario is transfer learning for few-shot image recognition on the Mini-ImageNet dataset [32], where the pretext task is image recognition on the base 64 classes, and the downstream task is image recognition on five novel classes with few labeled images per class, either 1-shot or 5-shot. For the base classes, we split their data into training and validation sets to evaluate base task performance. The experimental setting largely follows a recent work [4] for transfer learning.
Table 4: 5-way few-shot recognition on Mini-ImageNet.

| Methods     | Base classes | Novel classes |
|-------------|--------------|---------------|
|             | Acc          | 1-shot        | 5-shot        |
| Baseline [4]| 82.3         | 51.75 ± 0.80  | 74.27 ± 0.63  |
| Supervised  | 83.6         | 54.60 ± 0.80  | 74.50 ± 0.65  |
| MoCo-v2     | 75.3         | 52.14 ± 0.73  | 73.30 ± 0.59  |
| Exemplar-v2 | 79.9         | 55.33 ± 0.75  | 77.18 ± 0.61  |

Table 5: Facial landmark prediction on MAFL.

| Methods     | Landmark error |
|-------------|----------------|
| Scratch     | 24.6%          |
| Supervised  | 6.3%           |
| MoCo-v2     | 5.7%           |
| Exemplar-v2 | 5.8%           |

Figure 3: Visual results of transfer learning for facial landmark prediction.

The pretrained network learned from the base classes is fixed, and a linear classifier is finetuned for 100 rounds upon the output features for the novel classes.

As in our detection transfer study, we compare three pretraining methods: supervised cross entropy, unsupervised MoCo-v2 and supervised Exemplar-v2. Each method is trained with MoCo-v2 augmentations and optimized for 2000 epochs with a cosine learning rate decay scheduler for fair comparison. In finetuning the downstream task, since there exists much variance in the feature norms from different pretrained networks, we cross-validate the best learning rate for each method on the validation classes. Note that adding an additional batch normalization layer is problematic because as few as only 5 images (5-way 1-shot case) are available during finetuning.

We use the backbone network of ResNet18 [4] for the experiments. Results are shown in Table 4. Due to different optimizers and number of training epochs, our supervised pretraining protocol is stronger than the baseline protocol [4], leading to better results. The unsupervised pretraining method MoCo-v2 is weaker for both the base classes and the novel classes, suggesting that the pretraining task and the downstream task are well aligned semantically. Our exemplar-based approach obtains improvements over MoCo-v2 on the base classes, while outperforming the supervised baselines on the novel classes. This demonstrates that removing the explicit constraints on intra-class instances generalizes the model for better transfer learning on few-shot recognition.

4.2 Facial Landmark Prediction

We next consider the transfer learning scenario from face identification to facial landmark prediction on CelebA [19] and MAFL [39]. The pretext task is face identification on CelebA, and the downstream task is to predict 5 facial landmarks on the MAFL dataset. The facial landmark prediction is evaluated by the average euclidean distance of landmarks normalized by the inter-ocular distance.

As in the prior studies, we compare three pretraining methods: supervised cross entropy, unsupervised MoCo-v2 and supervised Exemplar-v2. Each method is trained with MoCo-v2 augmentations and optimized for 1400 epochs with a cosine learning rate decay scheduler. For landmark transfer, we finetune a two-layer network that maps the spatial output of ResNet50 features to landmark coordinates. The two-layer network contains a 1 × 1 convolutional layer that reduces 2048 channels to 128 and a fully connected layer, interleaved with LeakyReLU and batch normalization layers. We finetune all layers end-to-end for 200 epochs with a learning rate of 0.02 and a batch size of 128.

The experimental results are summarized in Table 5. Unsupervised pretraining by MoCo-v2 outperforms the supervised counterpart for this transfer, suggesting that the task misalignment between face identification and landmark prediction is large. In other words, faces corresponding to the same identity hardly reveal information about their poses. Our proposed exemplar-based pretraining approach weakens the influence of the pretraining semantics, leading to results that maintain the transfer performance of MoCo-v2. Qualitative results are displayed in Figure 3.
5 Related Works

Since the marriage of the ImageNet dataset [6] and deep neural networks [17], supervised ImageNet pretraining has proven to learn generic representations that facilitate a variety of applications such as high-level detection [11, 24] and segmentation [20], low-level texture synthesis [9], and style transfer [10]. ImageNet pretraining also works amazingly well under a large domain gap for medical imaging [22] and depth estimation [18]. The good transferability of ImageNet pretrained networks has been extensively studied [1, 2]. The transferability across each neural network layer has also been quantified for image classification [36], and a reduction of dataset size was found to have only a modest effect on transfer learning using AlexNet [15]. In addition, a correlation between ImageNet classification accuracy and transfer performance has been reported [10], and the benefit of ImageNet pretraining has been shown to become marginal when the target task has a sufficient amount of data [13].

Beyond ImageNet transfer, there has been an effort to discover the structures and relations among tasks for general transfer learning [25, 26]. Taskonomy [37] builds a relation graph over 22 visual tasks and systematically studies the task similarities. In [28], task cooperation and task competition are quantitatively measured to improve transfer learning. Similar phenomena are observed that task misalignment may lead to negative transfer [33] and that the number of layers that tasks may share depends on the task similarity [31]. Other works [8, 29] explore alternative methods to measure the structure and similarity between tasks.

While most works study transfer learning based on supervised pretraining, our work focuses on analyzing transfer learning based on unsupervised pretraining, particularly on MoCo [12] with the instance discrimination task [35]. Over recent years, the research community has achieved significant progress on unsupervised learning [35, 12, 3], closing the gap with supervised pretraining for transfer learning. However, little is understood about why instance discrimination pretraining leads to improved transfer learning performance. Our work is the first to shed light on this, and it uses this understanding to elevate the performance of supervised pretraining.

6 Conclusion

This work focuses on the downstream task of object detection on PASCAL VOC to analyze and better understand unsupervised pretraining. Through a fair protocol for comparing pretraining methods, we confirm the advantage of unsupervised pretraining. Our main findings that help to understand this transfer are as follows:

- When finetuning the network over all layers end-to-end with sufficient data, it is mainly low-level and mid-level representations that are transferred, not high-level representations. This suggests that unsupervised representations learned on various datasets share a common low-level and mid-level representation which performs similarly and adapts quickly towards the target problem.

- The output features from unsupervised pretrained networks are not agnostic to the pretrained datasets, as they are still overfit to the high-level semantics of the dataset being trained on.

- Unsupervised pretrained networks learned by the instance discrimination pretext task contain rich information for reconstructing pixels from the output features. In order to discriminate among all instances, the network appears to learn a holistic encoding over the entire image.

- For supervised pretraining, the intra-class invariance encourages the network to focus on discriminative patterns and disregards patterns uninformative for classification. This may lose information which could be useful for the target task when there is task misalignment. An exemplar-SVM style pretraining scheme based on the instance discrimination framework is shown to improve generalization for downstream applications.

While most insights and conclusions drawn above are taken from the single case of ImageNet to VOC detection transfer, we expect that these findings may have broad implications on other transfers. Our experiments on two other transfer scenarios confirm the generalization ability of exemplar-based pretraining. It is our hope that the presented studies provide insights that inspire better pretraining methods for transfer learning.
References

[1] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In European conference on computer vision, pages 329–344. Springer, 2014.

[2] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. From generic to specific deep representations for visual recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 36–45, 2015.

[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] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In International Conference on Learning Representations, 2019.

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

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[7] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE transactions on pattern analysis and machine intelligence, 2015.

[8] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12387–12396, 2019.

[9] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In Advances in neural information processing systems, pages 262–270, 2015.

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2414–2423, 2016.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[13] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In Proceedings of the IEEE International Conference on Computer Vision, pages 4918–4927, 2019.

[14] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In European conference on computer vision, pages 340–353. Springer, 2012.

[15] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? arXiv preprint arXiv:1608.08614, 2016.

[16] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2661–2671, 2019.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 2012.

[18] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5162–5170, 2015.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15:2018, 2018.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.
[21] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In 2011 International conference on computer vision, pages 89–96. IEEE, 2011.

[22] Romain Mormont, Pierre Geurts, and Raphaël Marée. Comparison of deep transfer learning strategies for digital pathology. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 2262–2271, 2018.

[23] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3234–3243, 2016.

[24] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.

[25] Daniel L Silver and Kristin P Bennett. Guest editor’s introduction: special issue on inductive transfer learning. Machine Learning, 73(3):215, 2008.

[26] Daniel L Silver, Qiang Yang, and Lianghao Li. Lifelong machine learning systems: Beyond learning algorithms. In 2013 AAAI spring symposium series, 2013.

[27] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In International Conference on Computer Vision (ICCV), 2017.

[28] Trevor Standley, Amir R Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? arXiv preprint arXiv:1905.07553, 2019.

[29] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In Proceedings of the IEEE International Conference on Computer Vision, pages 1395–1405, 2019.

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9446–9454, 2018.

[31] Simon Vandenhende, Stamatios Georgoulis, Bert De Brabandere, and Luc Van Gool. Branched multi-task networks: deciding what layers to share. arXiv preprint arXiv:1904.02920, 2019.

[32] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In Advances in neural information processing systems, pages 3630–3638, 2016.

[33] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 11293–11302, 2019.

[34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[35] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[36] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Advances in neural information processing systems, pages 3320–3328, 2014.

[37] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3712–3722, 2018.

[38] 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, pages 586–595, 2018.

[39] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In European conference on computer vision, pages 94–108. Springer, 2014.

[40] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. arXiv preprint arXiv:2004.06638, 2020.
A1 More Pretraining Epochs for Supervised Method

In the main text, 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 6, 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\textsubscript{75}, while it outperforms MoCo on AP\textsubscript{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.

Table 6: Longer supervised pretraining for object detection transfer on PASCAL VOC.

| Pretraining Epochs | ImageNet | VOC07 detection | VOC0712 detection |
|--------------------|----------|----------------|-------------------|
|                    | Acc | AP | AP\textsubscript{50} | AP\textsubscript{75} | AP | AP\textsubscript{50} | AP\textsubscript{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 |

A2 Additional Detection Results on VOC07+12

We provide additional object detection transfer results on VOC0712 for pretraining models on various datasets. In Table 7, these results align with the conclusions we draw from the VOC07 results. However, the advantages of unsupervised pretraining are smaller due to the larger transfer dataset for finetuning. We also provide a visualization of various datasets for training these models in Figure 4.

Table 7: Transfer performance on VOC0712 from pretraining on various datasets. “ImageNet (10%)” denotes subsampling 1/10 of images per class in 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 CelebA and Places models are trained with identity and scene categorization supervisions, and supervised COCO model is trained with bounding box supervision for object detection.

| Pretraining Data | #Imgs | Annotation | Supervised ImageNet | Supervised VOC07 detection | Supervised VOC0712 detection | Unsupervised ImageNet | Unsupervised VOC0712 detection |
|------------------|-------|------------|---------------------|---------------------------|-----------------------------|-----------------------|-----------------------------|
|                  |       |            | Acc | AP | AP\textsubscript{50} | AP\textsubscript{75} | Acc | AP | AP\textsubscript{50} | AP\textsubscript{75} |
| ImageNet         | 128K  | object     | 77.3 | 55.3 | 82.3 | 61.3 | 67.5 | 56.9 | 82.2 | 63.5 |
| ImageNet (10%)   | 128K  | object     | 57.8 | 53.3 | 80.8 | 58.6 | 58.9 | 55.6 | 81.4 | 62.0 |
| ImageNet-100     | 124K  | object     | 50.9 | 51.9 | 79.3 | 57.1 | 56.5 | 55.2 | 81.4 | 60.4 |
| Places           | 244K  | scene      | 52.3 | 51.2 | 79.2 | 55.4 | 57.1 | 56.7 | 81.7 | 63.4 |
| CelebA           | 163K  | identity   | 30.3 | 50.9 | 77.4 | 55.5 | 40.1 | 55.0 | 80.4 | 60.9 |
| COCO             | 118K  | bbox       | 57.8 | 58.5 | 83.0 | 65.6 | 50.6 | 55.5 | 81.5 | 62.1 |
| Synthia          | 365K  | segment    | 30.2 | 52.6 | 79.9 | 57.6 | 13.5 | 50.8 | 77.4 | 55.2 |

A3 Details on Image Reconstruction by Inverting Features

A3.1 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 \( C_{m}^{k} \) denote a Convolution-BatchNorm-LeakyReLU layer with \( k \) channels and \( m \times m \) spatial filters; \( CD_{m}^{k} \) denote a Convolution-Downsample-BatchNorm-LeakyReLU layer, and \( CU_{m}^{k} \) denote a Convolution-BatchNorm-LeakyReLU-Upsample layer. We use a stride of 2 for both the upsampling and down-sampling layers.

Encoder: \( CD_{16}^{7} - C_{16}^{7} - CD_{32}^{7} - C_{32}^{7} - CD_{64}^{7} - C_{64}^{7} - CD_{128}^{7} - C_{128}^{7} - CD_{128}^{7} - C_{128}^{7} - CD_{128}^{7} - C_{128}^{7} - CD_{128}^{7} - C_{128}^{7} \)

Decoder: \( C_{16}^{7} - CU_{16}^{7} - C_{32}^{7} - CU_{32}^{7} - C_{64}^{7} - CU_{64}^{7} - C_{128}^{7} - CU_{128}^{7} - C_{128}^{7} - CU_{128}^{7} - C_{128}^{7} - CU_{128}^{7} - C_{128}^{7} - CU_{128}^{7} \)
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.

**A3.2 More Visual Results**

We show more results on image reconstruction in Figure 5.

**A3.3 Evaluating Reconstructions by Perceptual Metrics**

Besides the averaged perceptual distances, we also provide a scatter plot of perceptual distance from individual reconstructions. In Figure 6, we can see that the reconstructions generated by MoCo are generally closer to the original images than those generated by the supervised method.

**A4 Ablation Study for Exemplar-based Pretraining**

Since our Exemplar pretraining uses a different set of parameters from MoCo, we provide an ablation study over the parameter $k$ and $\tau$ for ImageNet linear readout in Table 8.

**A5 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 7, 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.
Figure 5: Additional results of image reconstructions from pretrained networks.

Figure 6: Perceptual distance between the reconstruction and the original image on 1000 validation images.

Table 8: An ablation study of parameter $k$ and $\tau$ for MoCo and Exemplar pretraining.

| Methods    | $k$   | $\tau$ | 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         |
Figure 7: Distribution of four types of false positives for each category.