Tradeoffs Between Contrastive and Supervised Learning: An Empirical Study

Ananya Karthik, Mike Wu, Noah Goodman, Alex Tamkin
Department of Computer Science
Stanford University
{ananya23,wumike,ngoodman,atamkin}@stanford.edu

Abstract

Contrastive learning has made considerable progress in computer vision, outperforming supervised pretraining on a range of downstream datasets. However, is contrastive learning the better choice in all situations? We show it is not. First, under sufficiently small pretraining budgets, supervised pretraining on ImageNet consistently outperforms a comparable contrastive model on eight diverse image classification datasets. This suggests that the common practice of comparing pretraining approaches at hundreds or thousands of epochs may not produce actionable insights for those with more limited compute budgets. Second, even with larger pretraining budgets we identify tasks where supervised learning prevails, perhaps because the object-centric bias of supervised pretraining makes the model more resilient to common corruptions and spurious foreground-background correlations. These results underscore the need to characterize tradeoffs of different pretraining objectives across a wider range of contexts and training regimes.

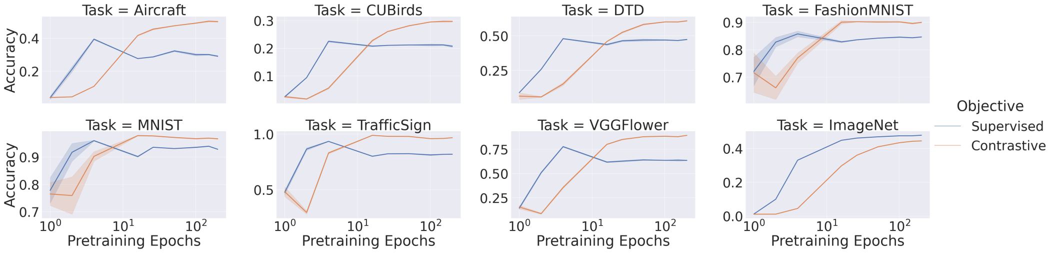


Figure 1: **Downstream accuracy of contrastive and supervised models for different pretraining budgets.** Models pretrained on ImageNet, then evaluated on 8 diverse image classification datasets. Shaded regions show the standard deviation across three runs (often too small to see without magnification; shown for all except ImageNet which had only one trial). Unpretrained models shown on far left of each plot.

1 Introduction

The cost of labeling large-scale datasets has motivated a rise in self-supervised pretraining, with recent methods in computer vision closing the gap with or even surpassing supervised approaches [1, 9, 11, 2, 32]. Instead of using labels, recent contrastive learning methods leverage an instance discrimination task [29, 6, 22, 2, 24] to achieve state-of-the-art results on a variety of computer vision tasks. The instance discrimination task treats each image as its own class, training a model

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

to determine whether two augmented examples were derived from the same original instance using a contrastive loss [10, 12, 27]. This training procedure produces models whose representations are broadly useful for a range of transfer tasks [7, 31].

These advances underscore the need for studying the real-world tradeoffs between contrastive and supervised pretraining. We investigate by comparing the transfer performance of both methods across different pretraining budgets and transfer datasets. Our results address two questions:

1. **Is contrastive learning better than supervised across all compute budgets?** No, different pretraining algorithms produce better representations at different pretraining budgets. Moreover, transfer accuracy on different tasks is not even monotonic across pretraining. Thus, we recommend that future work on pretraining report transfer accuracy across epochs so practitioners can make informed decisions based on their end task and compute budget.
2. **For larger compute budgets, is contrastive pretraining better for all tasks?** No. While the supervised model eventually achieves worse downstream accuracy than the contrastive model on most tasks, we identify tasks where the object-centric bias of ImageNet pretraining aids transfer—especially in the Waterbirds dataset, which measures reliance on spurious correlations and ImageNet-C, which measures robustness to common corruptions.

2 Related Work

Performance of self-supervised learning Previous studies on representation learning for visual tasks have provided insights into the generalizability and transfer performance of various algorithms, including the comparison of supervised and unsupervised learning methods [4, 18, 17, 8, 16, 33, 7, 31]. In particular, Ericsson et al. [7] find that the best self-supervised models outperform a supervised baseline on most datasets in their benchmark. Our work builds on this analysis by holding variables like pretraining epochs and data augmentations constant, performing a controlled analysis of these models' learning dynamics and transferability.

Sample efficiency of pretraining methods Several studies [33, 8, 16, 31] analyze sample efficiency or computational efficiency during transfer or finetuning, including Zhao et al. [34], which compares the pretraining dynamics of supervised and unsupervised learning methods on the VOC'07 detection task. However, to the best of our knowledge, there has not been a comprehensive comparison of the learning dynamics over the course of pretraining time for both supervised and contrastive learning across a diverse range of downstream image classification tasks.

How pretraining objectives shape model representations Zhao et al. [34] visualize the representations of contrastive and supervised models, arguing that the former objective may produce more holistic representations compared to the latter. Furthermore, Ericsson et al. [7] observe that self-supervised pretraining attends to larger regions than supervised pretraining, a characteristic that may aid the transfer performance of self-supervised methods. Cole et al. [4] and Horn et al. [14] show that self-supervised pretraining does not outperform supervised learning for fine-grained classification tasks. Our study builds upon these works by providing examples of specific object-centric *tasks* where the supervised model achieves higher accuracy, as well as cases where the holistic representations of the contrastive model prevail.

3 Experiments

3.1 Experimental Settings

We pretrain two ResNet-18 models on ILSVRC-2012 (ImageNet) [25] for 200 epochs with a batch size of 128. We use the standard cross entropy loss for the supervised model, and we use the InfoNCE objective from Wu et al. [29] for the contrastive model, leveraging a memory bank for negatives. Both models are pretrained with identical image augmentations, the same as Chen et al. [2] without random Gaussian blur, and identical model architecture. For pretraining, we use SGD with a learning rate of 0.03, momentum of 0.9, and weight decay of 1e-4.

For transfer, we use the linear evaluation protocol [2], training a logistic regression model on the outputs of the prepool 512x7x7 layer of a frozen pretrained model. We evaluate both pretrained models by training them for 100 epochs on eight transfer tasks: MNIST [19], FashionMNIST [30], VGG Flower (VGGFlower) [23], Traffic Signs (TrafficSign) [15], Describable Textures (DTD) [3], CUB-200-2011 (CUBirds) [28], Aircraft [21], and ImageNet itself. (See Table 1 in the Appendix.) We use SGD with a batch size of 256, learning rate of 0.01, momentum of 0.9, and weight decay of 1e-4.

3.2 Results

We first compare final transfer accuracies achieved by contrastive and supervised pretraining (Figure 4a). In line with previous studies [7, 2], we find that the contrastive model outperforms the supervised model on all transfer tasks except ImageNet, its pretraining dataset.

3.2.1 Learning Dynamics Across Compute Budgets

We also investigate the representation learning dynamics and computational efficiency of the two models. Transfer accuracy by pretraining budget is shown in Figure 1. All results except ImageNet represent the average of three trials with different random seeds; ImageNet results represent one trial. We observe the following trends across the seven non-ImageNet tasks:

(i) With only a few epochs of pretraining, the supervised model maintains a lead over the contrastive model. However, by 15 epochs the contrastive model rapidly overtakes the supervised model's downstream accuracy, maintaining a lead until the end of pretraining. Thus, the contrastive model is more computationally efficient for all but the most restricted compute budgets. However, it also suggests a note of caution: models that prevail after a certain number of pretraining steps may not always win out at other, more modest budgets.

(ii) The downstream accuracy of both models does not always increase monotonically across pretraining. Particularly pronounced for the supervised model, this phenomenon suggests a potential misalignment between the representations developed for the supervised task and those most useful for the downstream tasks.

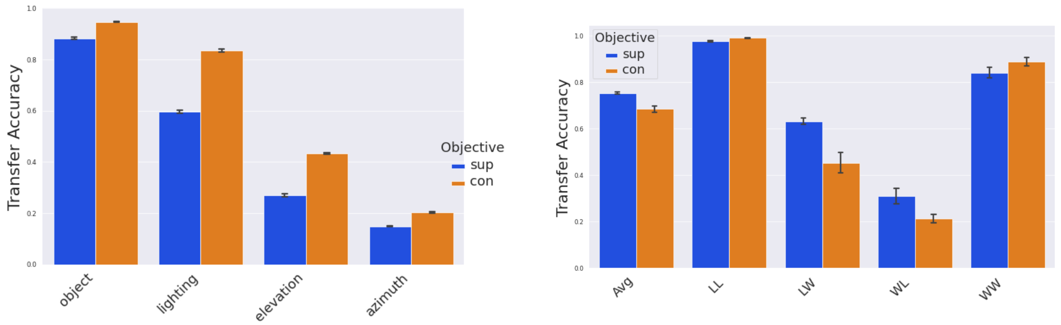
3.2.2 Downstream Effects of Biases Acquired During Pretraining

We have demonstrated that the supervised and contrastive models have different pretraining dynamics, suggesting that the models may acquire different feature processing capabilities during pretraining. But what are the downstream effects of these representational differences? Zhao et al. [34] conclude that supervised pretraining may learn more object-specific features than contrastive models. In three controlled studies (see Figure 3 in the Appendix for example images from the datasets used), we investigate this hypothesis by examining specific tasks where an object-centric bias may be salient. All results are averages of three trials with different random seeds.

NORB We study the transfer performance of both models on a carefully-controlled dataset which isolates the models' abilities to capture both object and non-object information in their representations. Specifically, we use NORB (small set[1]) — synthetic images of 50 types of toys, annotated with toy category, lighting conditions, elevations, and azimuths [20]. Contrastive pretraining outperforms supervised learning on object, elevation, lighting, and azimuth classification tasks. For the non-object elevation and lighting transformations, the gap in accuracy between the models was pronounced — 16.18% and 23.91%, respectively — possibly due to the supervised model developing more object-centric representations. However, drawing firm conclusions is challenging, as the accuracy difference across tasks may be misleading when object classification accuracy approaches the 100% ceiling. Furthermore, even if the gap differs, the contrastive model still outperforms the supervised model across tasks. Thus, these results provide relatively modest evidence of transfer tasks where object-centricity impacts the two models differently.

Waterbirds We then expand our experiments to a different set of non-object properties — the content of image backgrounds. We evaluate contrastive and supervised pretraining on Waterbirds, a dataset designed to examine spurious correlations based on the relationship between object and background (see Appendix) [26]. We find that on images in which the backgrounds and objects are

[1] `https://cs.nyu.edu/~ylclab/data/norb-v1.0-small/`



(a) **Transfer accuracy on NORB object, elevation, lighting, and azimuth classification.** Contrastive accuracy was higher than supervised accuracy on all tasks. Models pretrained on ImageNet for 200 epochs; average of three trials, with error bars for standard deviations.

(b) **On the Waterbirds dataset, the supervised model appears to attend less to spurious correlations. LW** indicates images of **L**and birds on **W**ater backgrounds. Models pretrained on ImageNet for 200 epochs; average of three trials, with error bars for standard deviations.

Figure 2: **Difference in learned representations: NORB and Waterbirds.**

mismatched the supervised model achieves higher transfer accuracy than the contrastive model, in contrast to the previous results showing higher image classification performance for the contrastive model. This suggests that the supervised model may have learned more object- or foreground-centric representations, which render the spurious background feature less prominent. While this result lends support to the notion that the contrastive model learns a more holistic image representation, it also suggests that the inductive bias attained from a more tailored representation may be helpful in underspecified settings where undesired features are also predictive of the class of interest [5].

ImageNet-C Last, we study the degradation of transfer performance in the presence of non-object-based corruptions. We hypothesized that if supervised learning results in more object-centric representations, then transfer performance might degrade less with non-object corruptions such as color shifts and changes in contrast. We evaluate both models, after transfer was performed for ImageNet, on 15 corruptions from ImageNet-C, a dataset created by applying 15 corruptions at 5 severity levels to ImageNet validation images [13]. We observe the relative mCE, which measures the performance degradation from clean to corrupted data (lower is better), to be lower for supervised ($\mathbf{91.08 \pm 0.279\%}$) vs the contrastive model ($\mathbf{95.41 \pm 0.157\%}$). This provides additional evidence that supervised pretraining may lead to more object-centric representations than contrastive approaches.

4 Discussion

We investigate tradeoffs between supervised and contrastive pretraining.

Our first set of experiments examines how the linear evaluation performance on a range of transfer tasks changes as each model pretrains. Surprisingly, we find that transfer performance does not monotonically increase across pretraining, suggesting a misalignment between representations learned for pretraining vs transfer. Moreover, while the contrastive model eventually achieves higher performance, for the first 10-15 epochs the supervised model yields better representations for downstream tasks. This not only reveals differences in the process by which both models acquire their useful representations, but also that conclusions drawn for models trained for thousands of epochs may not always transfer over to practitioners with more modest compute budgets. Thus, we encourage developers of new pretraining techniques to release learning dynamics curves so that practitioners can make decisions based on their own budgets and use cases.

To further explore tradeoffs between the two models, we examine whether supervised learning imparts an object-centric bias detectable through improved performance on transfer tasks. We find strong effects in the case of Waterbirds and ImageNet-C, but weaker effects for the NORB dataset. We encourage future work investigating how pretraining objectives shape the behavior of models in ambiguous scenarios, as well as more broadly investigating whether these conclusions hold across a wider range of architectures, hyperparameters, datasets, and training objectives.

References

[1] Mathilde Caron, Ishan Misra, J. Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *ArXiv*, abs/2006.09882, 2020.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

[3] M. Cimpoi, Subhransu Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[4] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work?, 2021.

[5] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[6] A. Dosovitskiy, P. Fischer, Jost Tobias Springenberg, Martin A. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1734–1747, 2016.

[7] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? *ArXiv*, abs/2011.13377, 2020.

[8] Priya Goyal, D. Mahajan, A. Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6390–6399, 2019.

[9] Jean-Bastien Grill, Florian Strub, Florent Altch'e, C. Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, B. A. Pires, Zhaohan Daniel Guo, M. G. Azar, Bilal Piot, K. Kavukcuoglu, R. Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv*, abs/2006.07733, 2020.

[10] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[12] Olivier J. Hénaff, A. Srinivas, J. Fauw, Ali Razavi, Carl Doersch, S. Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. *ArXiv*, abs/1905.09272, 2020.

[13] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1903.12261, 2019.

[14] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections, 2021.

[15] Sebastian Houben, J. Stallkamp, J. Salmen, Marc Schlipsing, and C. Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013.

[16] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1920–1929, 2019.

[17] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2666, 2019.

[18] Klemen Kotar, Gabriel Ilharco, Ludwig Schmidt, Kiana Ehsani, and Roozbeh Mottaghi. Contrasting contrastive self-supervised representation learning pipelines, 2021.

[19] Y. LeCun, L. Bottou, Yoshua Bengio, and P. Haffner. Gradient-based learning applied to document recognition. 1998.

[20] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2:II–104 Vol.2, 2004.

[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, abs/1306.5151, 2013.

[22] Ishan Misra and L. V. D. Maaten. Self-supervised learning of pretext-invariant representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716, 2020.

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

[24] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *ArXiv*, abs/2007.13916, 2020.

[25] Olga Russakovsky, J. Deng, Hao Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[26] Shiori Sagawa, Pang Wei Koh, T. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv*, abs/1911.08731, 2019.

[27] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.

[28] C. Wah, Steve Branson, P. Welinder, P. Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[29] Zhirong Wu, Yuanjun Xiong, S. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *ArXiv*, abs/1805.01978, 2018.

[30] H. Xiao, K. Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017.

[31] Xingyi Yang, Xuehai He, Yuxiao Liang, Yue Yang, Shanghang Zhang, and P. Xie. Transfer learning or self-supervised learning? a tale of two pretraining paradigms. *ArXiv*, abs/2007.04234, 2020.

[32] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

[33] Xiaohua Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, A. Dosovitskiy, Lucas Beyer, Olivier Bachem, M. Tschannen, Marcin Michalski, O. Bousquet, S. Gelly, and N. Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[34] Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *ArXiv*, abs/2006.06606, 2020.

Table 1: **Details of datasets used.**

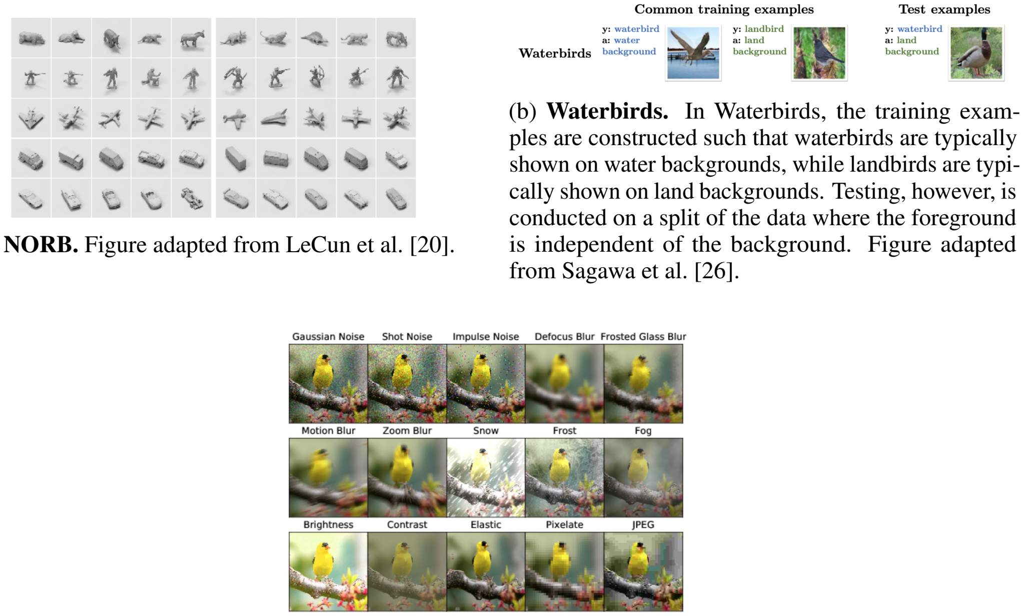
Dataset	Focus	Classes	Train Size
VGG Flower	flowers	102	6507
Traffic Sign	road signs	43	31367
MNIST	digits	10	60000
Fashion MNIST	apparel	10	60000
DTD	textures	47	3760
CU Birds	birds	200	5994
Aircraft	aircrafts	100	3334
ImageNet	diverse	1000	1281167
Norb-Object	toys	6	48600
Norb-Elevation	elevations	9	48600
Norb-Lighting	lighting	6	48600
Norb-Azimuth	rotations	18	48600
Waterbirds	birds	2	4795

A Appendix

A.1 Details of Datasets Used

In Table 1, we show the details of the 10 datasets used in this study, including the focus, number of classes, and the size of the training set.

Example images from the datasets used in Section 3.2.2 are shown in Figure 3. Each represents a different way of examining the object-centricity of a model's representations.



(a) **NORB.** Figure adapted from LeCun et al. [20].

(b) **Waterbirds.** In Waterbirds, the training examples are constructed such that waterbirds are typically shown on water backgrounds, while landbirds are typically shown on land backgrounds. Testing, however, is conducted on a split of the data where the foreground is independent of the background. Figure adapted from Sagawa et al. [26].

(c) **ImageNet-C.** Figure adapted from Hendrycks and Dietterich [13].

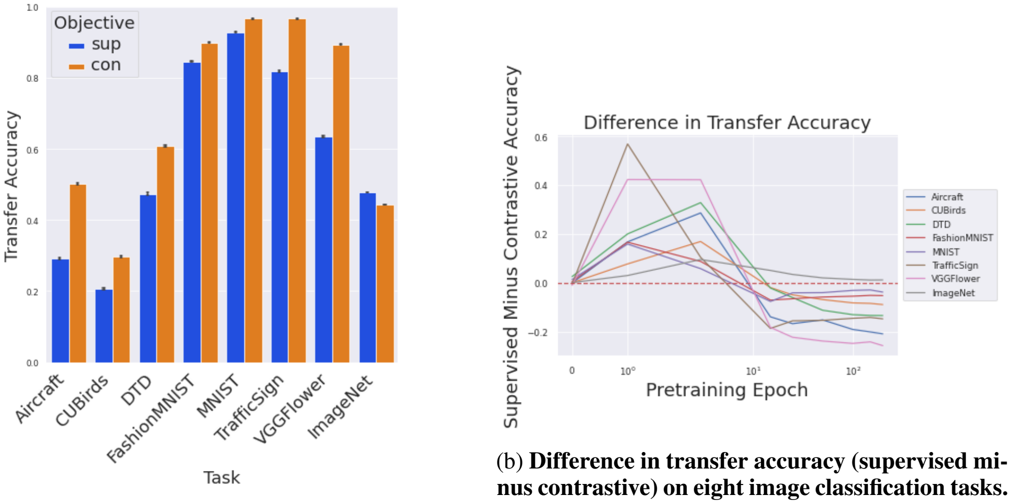
Figure 3: **Example images from datasets used in Section 3.2.2.**

A.2 Transfer Accuracies on Diverse Tasks

In Table 2, we compare the transfer accuracies on 8 diverse tasks after 200 epochs of pretraining. We visualize this comparison in Figure 4a, and we explore the difference in transfer accuracy between the supervised and contrastive models over the course of pretraining in Figure 4b.

Table 2: **Comparison of transfer accuracies on diverse tasks.** After 200 epochs, the contrastive model achieves higher transfer accuracy for all tasks except ImageNet, which was used to pretrain the supervised model. Values after ± are standard deviations.

Task	Supervised	Contrastive
Aircraft	29.1 ± 0.3	**50.0** ± 0.3
CUBirds	20.7 ± 0.3	**29.7** ± 0.2
FashionMNIST	84.6 ± 0.1	**89.9** ± 0.1
DTD	47.4 ± 0.3	**60.8** ± 0.2
TrafficSign	81.8 ± 0.2	**96.6** ± 0.1
MNIST	92.8 ± 0.1	**96.7** ± 0.1
VGGFlower	63.6 ± 0.2	**89.4** ± 0.1
ImageNet	**47.8** ± 0.0	44.4 ± 0.0



(a) **Comparison of transfer accuracies achieved by supervised and contrastive pretraining across 8 diverse image classification transfer tasks.** Both models were pretrained on ImageNet for 200 epochs. Contrastive accuracy was higher than supervised accuracy on all tasks except ImageNet. Results on all tasks represent the average of three independent runs, with error bars representing the standard deviation.

(b) **Difference in transfer accuracy (supervised minus contrastive) on eight image classification tasks.** The dashed red line indicates when contrastive and supervised accuracies match, and we see that every task except ImageNet crosses the dashed line from positive to negative—indicating that contrastive accuracy overtakes supervised accuracy—at or before epoch 15 of pretraining.

Figure 4: **Transfer accuracies on diverse tasks.**