Perceptual Inductive Bias Is What You Need Before Contrastive Learning

Supplementary Material

This supplementary material provides: (1) complete implementation details for contrastive pretraining, linear classification, transfer learning, segmentation, and depth estimation; and (2) additional results and analyses, including performance on the STL-10 dataset, convergence curves of MidVCL and hybrid MidVCL, ablation studies, and a preliminary scalability analysis.

1. Implementation details

Pretraining We train all baselines from scratch on ImageNet-100 and STL-10 using the optimizer and hyperparameters reported on solo-learn library [3]. For our proposed approaches, during the contrastive pretraining phase, the model is warmed-up in the first 20 epochs in the same way as stated in PCL [2]. We set $\tau = 0.1, \alpha = 10$, and the number of clusters $K = \{2000, 4000, 6000\}$ for S-PCL. We set the number of negative samples $r = 1024$ when training on ImageNet-100 and STL-10. For hybrid frameworks, the first 100 epochs are trained using either S-PCL or MidVCL and are trained using either PCL or MoCoV2. We using a cosine learning rate scheduler in the pretraining. For the MidVCL, we set alpha and beta to both 0.5. In S-PCL, ReflCL, ShadCL, and MidVCL, data augmentations are only applied to original images, not to shape silhouette or intrinsic images.

Linear classification We use the entire labeled training set of ImageNet-100 and STL10 to train the linear classifier on the fix representation. The linear classifier is trained for 100 epochs and the batch size is 256. The SGD optimizer is adopted, with the initial learning rate of 0.3 and momentum of 0.9. The learning rate will be multiplied by 0.1 at epoch 60 and 80.

Transfer learning to ImageNet-1K We use the entire labeled training set of ImageNet-1K to train the linear classifier on the fix representation. The linear classifier is trained for 20 epochs and the batch size is 256. The loss function is cross entropy. The SGD optimizer is adopted, with the initial learning rate of 5 and momentum of 0.9. The learning rate will be multiplied by 0.1 at epoch 12 and 16.

Segmentation A SegFormer segmentation head is added after the ResNet18 backbone. The image size is 256 by 256. The optimizer used is the AdamW with initial learning rate of 0.001 and weight decay 0.1. We use a polynomial learning rate scheduler. We fine-tune both the backbone and the segmentation head. For ADE20K, the model is trained for 200 epochs. For Cityscapes, the model is trained for 500 epochs.

Depth estimation Similar to the segmentation, we add a Feature Pyramid Network (FPN) head on the basis of ResNet18 and fine-tune both backbone and head using NYU Depth v2 dataset. The loss fucntion is L1 loss. The optimizer used is Adam with inital learning rate of 0.0001 and weight decay 0.02. The learing rate scheduler is OneCycleLR. The model is trained for 10 epochs.

2. Experiment on STL-10 dataset

After pretraining using the ssl subset, we trained a linear classifier on frozen representations using the entire labeled training set of STL-10, and here we report the Top-1 and Top-5 accuracies on the validation sets. As illustrated in Table.1. The trend is the same as that of the results in ImageNet-100 in the main paper.

3. Ablation studies

Combination of losses We carry out ablation studies on different combinations of loss functions on the STL10 dataset. Results are reported in Table.2. We observe that reflectance and shading alone would yield trivial solution (model collapse). However, combining contrastive learning loss (InfoNCE) with any of the perceptual losses would consistently yield better performance than InfoNCE loss or perceptual loss alone on classification task.

Training order of hybrid frameworks We also explore the effect of the order of hybrid sequential training frameworks. The experiments are conducted on STL10 dataset. The model is trained for 200 epochs in total. Within 200 epochs of training, 100 epochs are trained by S-PCL and the other 100 epochs are trained by either MoCoV2 or PCL. From Tabel.3, we observe that the sequence in which the

Table 1. Linear classification accuracies of different approaches trained for 100 and 400 epochs on STL-10 dataset. The best performances are in bold.

	Epochs	STL10 Top-1 Acc (%)↑	STL10 Top-5 Acc (%)↑
SimCLR	100	82.7	99.4
MoCoV2	100	81.6	99.3
PCL	100	80.8	99.2
S-PCL(Ours)	100	**87.5**	**99.7**
ReflCL(Ours)	100	82.0	98.8
ShadCL(Ours)	100	81.5	98.7
SimCLR	400	88.9	99.7
MoCoV2	400	87.9	99.7
PCL	400	88.0	99.5
S-PCL(Ours)	400	88.7	99.7
ReflCL(Ours)	400	88.2	99.6
ShadCL(Ours)	400	86.2	99.4
S-PCL+PCL(Ours)	400	**91.7**	**99.8**
S-PCL+MoCoV2(Ours)	400	91.5	**99.8**

Table 2. Linear classification accuracies of frameworks using standard contrastive loss and different perceptual losses in isolation and in combination.

STL10	Epochs	Top-1(%)↑	Top-5(%)↑
InfoNCE	200	81.6	99.3
Only Shape	200	75.7	98.3
InfoNCE+Shape	200	89.0	99.5
Only Refl	200	10.0	10.0
InfoNCE+Refl	200	83.3	99.0
Only Shad	200	10.0	10.0
InfoNCE+Shad	200	82.5	98.9

training frameworks are applied significantly influences the performance. Specifically, models that begin training with S-PCL first exhibit better performance. This suggests that the shape feature plays an important role in the early stage. It is beneficial to learn shape representations before semantic representations.

Table 3. Linear classification accuracies on STL-10 of hybrid frameworks with different training orders.

	Epochs Pretrained	STL10 Top-1 Acc (%)	STL10 Top-5 Acc (%)
MoCoV2+S-PCL	100+100	87.8	99.5
S-PCL+MoCoV2	100+100	**90.7**	**99.7**
PCL+S-PCL	100+100	88.0	**99.7**
S-PCL+PCL	100+100	**91.2**	**99.7**

Different offline silhouette generation methods Since the offline shape silhouette method we used, TRACER, is a supervised method, it may be the case that our performance gain is due to label leakage. To test this, we carry out the same experiment of S-PCL under the same hyperparameter setting using silhouettes generated by MOVE [1], an unsupervised method. We test on epoch 50, 100 and 150. The results and trends are almost the same, as shown in Figure.1, suggesting that the gains of our proposed method are not due to the label leakage.

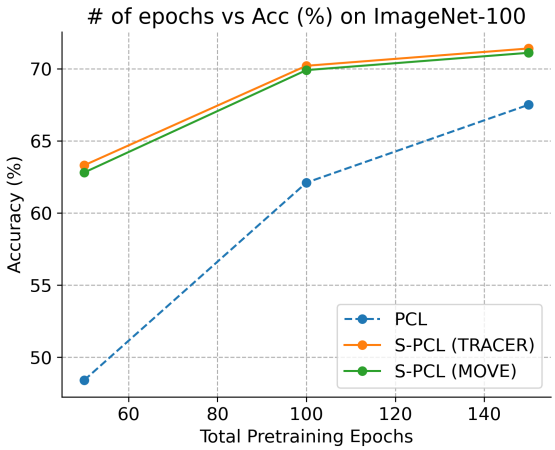

Figure 1. Linear classification accurcies in ImageNet-100 for S-PCL using different silhouette generation methods.

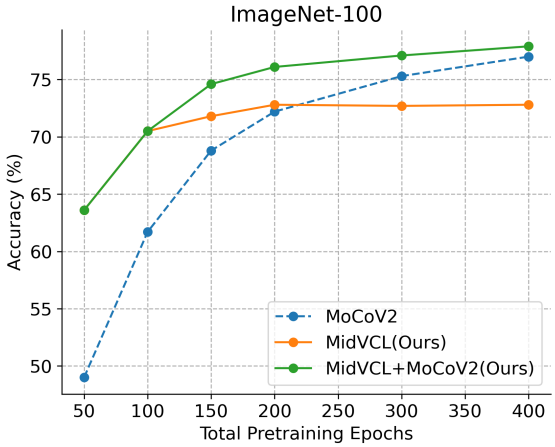

Figure 2. Linear classification accurcies on ImageNet-100 for MidVCL and hybrid MidVCL.

4. More analysis on ReflCL, ShadCL and MidVCL

Convergence curve of MidVCL The convergence curve of MidVCL is shown in Figure.2. It demonstrates the same trend as the S-PCL, accelerating the convergence of classification by 2x.

Sensitivity map analysis of ReflCL, ShadCL We perform sensitivity map analysis on ReflCL and ShadCL to study what features in the original image those differently trained models pay attention to, see Figure.3. Combining results in the main paper, we can find that for shape, we can learn that global shape clusters and use that as prior while this is not feasible for reflectance and shading.

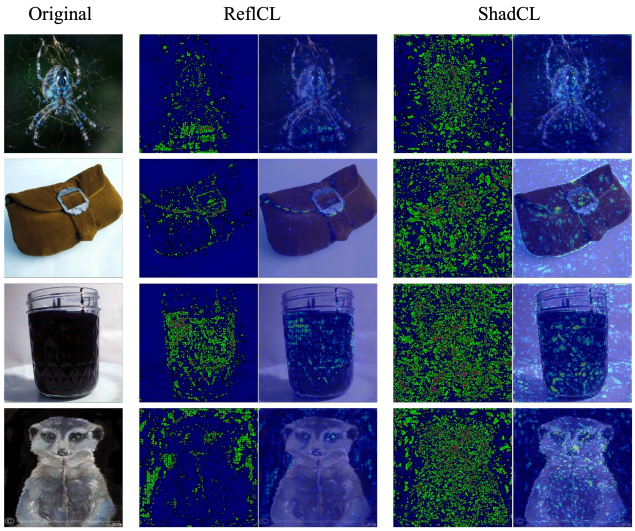

Figure 3. Image features highlighted by ReflCL and ShadCL respectively, as revealed by Smoothgrad sensitivity analysis.

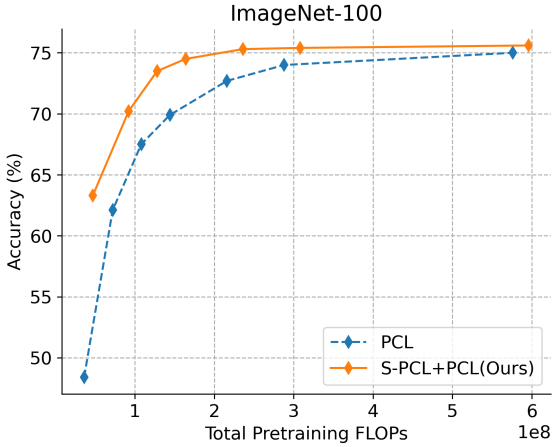

Figure 4. Linear classification accuracy in ImageNet-100 vs Total pretraining FLOPs. Offline shape mask generation is included for S-PCL+PCL(Ours).

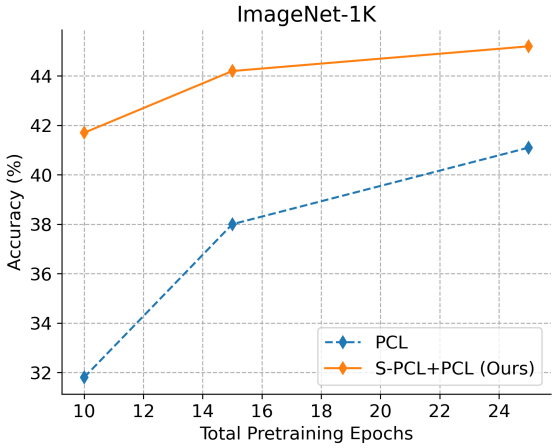

Figure 5. Initial results on ImageNet-1k.

5. Scalability

Total FLOPs estimation of S-PCL One may doubt the scaling behaviors of offline silhouette generation. To address this concern, we estimate the total computation (FLOPs) including offline preprocessing and online training for S-PCL. For IN100, the offline preprocessing requires much less computation compared with training (1.9e6 GFLOPs vs 3.0e8 GFLOPs), especially with large number of training epochs. The Acc vs total GFLOPs is shown in Figure.4. It demonstrates that our proposed approach can reach the same accuracy only with the half of the total computation even taking into account the offline generation.

Preliminary result on large model and datasets To address the concern of scaling behavior to large model and dataset, we conduct experiment on the ImageNet-1k dataset and using ResNet50. In the pretraining phase, the first 10 epochs are PCL and the remaining epochs are S-PCL. Then, we use all training set of ImageNet-1k to train a linear classifier on the frozen reprsentation. The linear classification accuracy of validation set is presented in the Figure.5. The curve shows a preliminary performance trend similar to those we demonstrated earlier on small datasets. We should emphasize that the primary focus of our work is on demonstrating the benefits of integrating the representation of mid-level visual representations, such as boundary shapes and intrinsic images, into self-supervised contrastive learning. It is possible that such benefits might dissipate when the dataset becomes very large. The fact that such benefits can be demonstrated for more modest datasets is important from the perspective of efficiency as well as biological plausibil-

ity. We are still running experiment in ImageNet-1k, and the results will be released in the future.

References

[1] Adam Bielski and Paolo Favaro. Move: Unsupervised movable object segmentation and detection, 2022. 2

[2] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 1

[3] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010. 1