Self-supervised Pre-training with Adaptive Augmentation and Equivariance-Coherent Representations Learning

Qin Wang[1] **Kai Krajsek**[2] **Hanno Scharr**[1]

[1] IAS-8: Data Analytics and Machine Learning, Forschungszentrum Jülich, Germany

[2] Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, Germany

{qi.wang, k.krajsek, h.scharr}@fz-juelich.de

Motivation

- Equivariance in feature learning ensures that a model's learned representations remain consistent under various transformations, including 2D or 3D translations, rotations, scaling, and changes in colour or illumination. This means that the information about the transformation can still be retrieved from the feature representation.
- Most Augmentation-based self-supervised (SSL) pretraining approaches foster invariant features, while equivariant features might be lost during the SSL learning process.
- Most Augmented-based SSL methods that take care on equivariant feature representation do not reach state-of-the art performance on competitive benchmarks or restrict their evaluation on synthetic data (e.g. SIE [2]). This limitation is the original motivation for this study.

Contribution

- Reconstruction as an auxilary task to learn equivariance.
- We demonstrate the effectiveness of our method on both artificial (3DIEBench [2]) and natural (ImageNet [3]) datasets, showing comparable (3DIEBench [2]) and improved performance (ImageNet) compared to existing baselines.
- Extensive evaluations on various downstream tasks, such as classification tasks, dense prediction tasks, homography estimation.

Computation cost

Backbone	N umber of GPUs	time(s)/epoch	training epochs
Pretraining procedure			
ViT-Small	8	600	800
ViT-Large	32	2900	150
Linear probing, average on all datasets			
ViT-Small	4	22	20
ViT-Large	8	60	20
Finetuning, average on all datasets			
ViT-Small	4	540	50
ViT-Large	8	2100	50

- We pretrain our model using ViT-Small and ViT-Large backbones. For hyperparameter search, each configuration is run more than 50 times.
- Linear probing and fine-tuning vary by task—COCO takes 30× longer to train than CIFAR-10.

References

[1] Maurice Weiler, Patrick Forr´e, Erik Verlinde, and Max Welling, Equivariant and Coordinate Inde-pendent Convolutional Networks, 2023.

[2] Q. Garrido, L. Najman, and Y. LeCun, "Self-supervised learning of split invariant equivariant representations," in Proceedings of the 40th International Conference on Machine Learning (ICML), 2023

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015.

[4] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "iBOT: Image BERT pre-training with online tokenizer," International Conference on Learning Representations (ICLR), 2022.

[5] Jülich Supercomputing Centre. (2021). JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. Journal of large-scale research facilities, 7, A183. http://dx.doi.org/10.17815/jlsrf-7-183

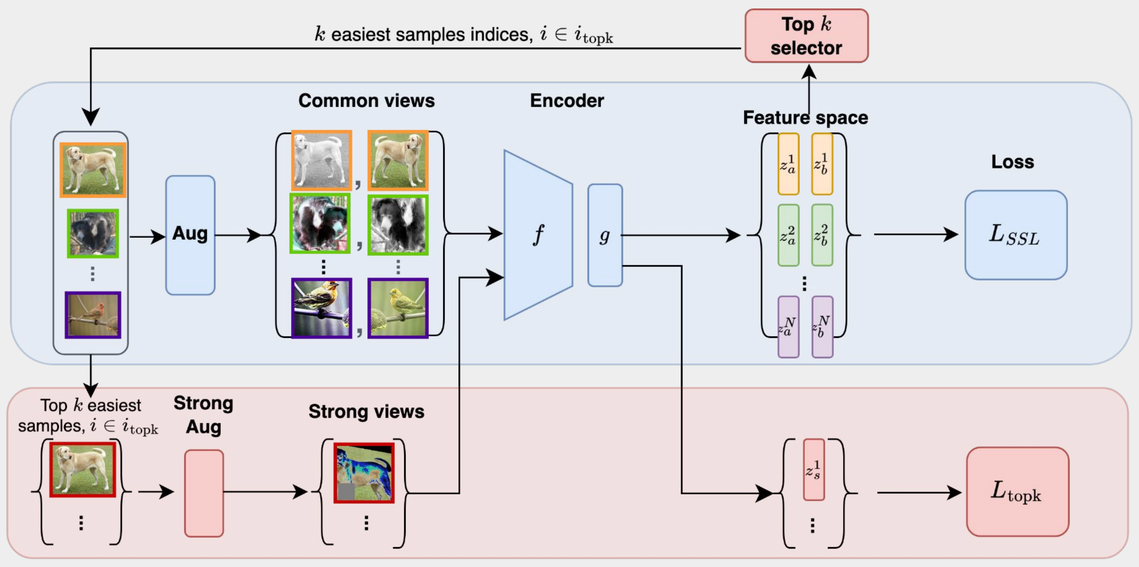
[6] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. 2018. Unified Perceptual Parsing for Scene Understanding. In Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part V. Springer-Verlag, Berlin, Heidelberg, 432–448.

[7] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

Acknowledgement

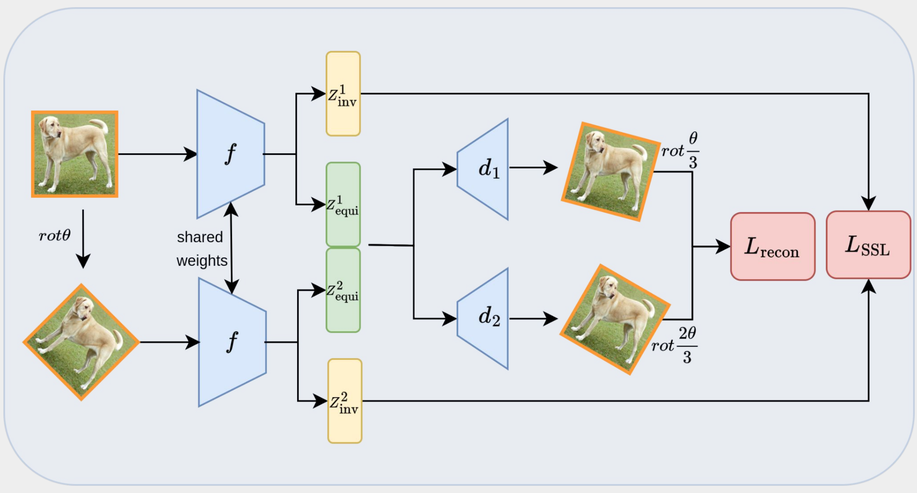
The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS [5] at Jülich Supercomputing Centre (JSC).

Adaptive augmentation for rescuing easy samples



- The **easiest samples** are selected based on the top-k highest cosine similarities. Strong augmentations are then applied to these samples to generate **strong views**.
- These strong views are fed into the model to compute the top-k loss, which is combined linearly with the original self-supervised learning (SSL) loss.

Equivariance-coherent representation learning



- Equivariant features are concatenated as input into different decoders **d** to reconstruct intermediate transformed images.
- Invariant features are matched with state-of-the-art SSL methods, such as iBOT [4].

Transformation estimations

Evaluation on 3DIEBench [2] dataset.

3DIEBench	Classification	Rotation Prediction	Color Prediction
SIE(rot)	**0.820**	**0.724**	0.054
SIE(rot+color)	0.809	0.502	**0.980**
Ours	0.782	0.554	0.954

Evaluation on ImageNet [3].

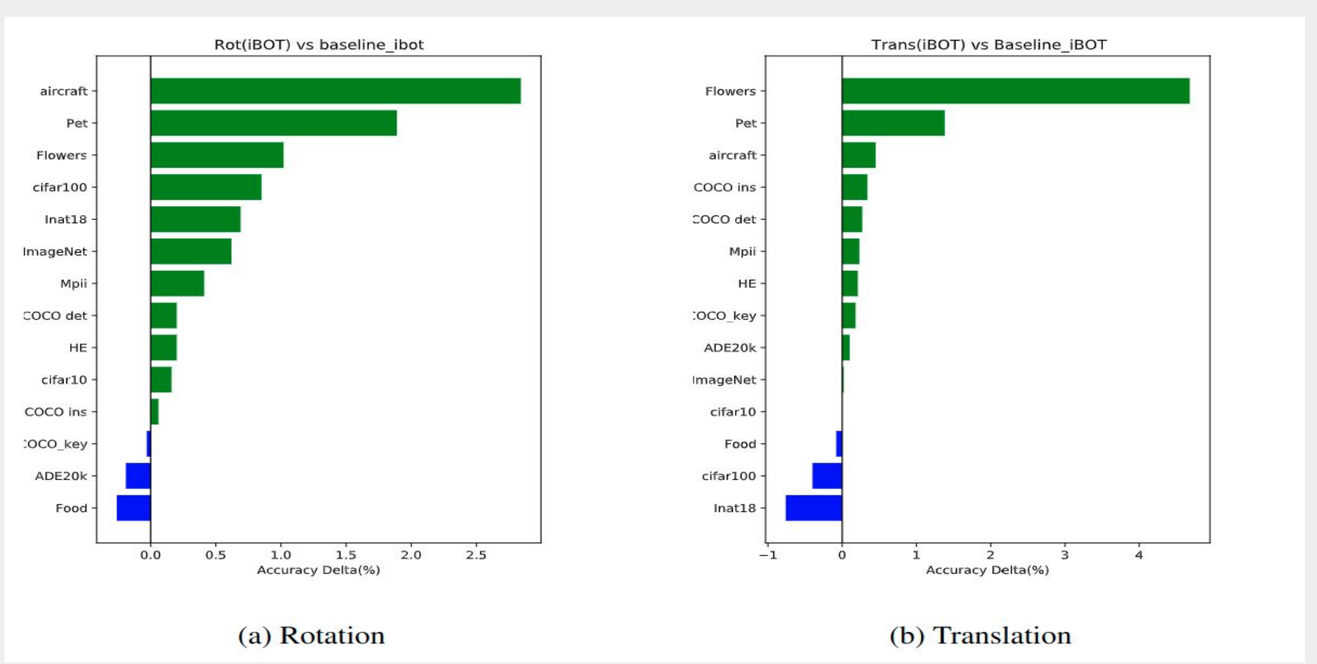
Metrics	R^2(rot)	R^2(color)	R^2(blur)	R^2(trans)
SIE(rot)	**0.990**	0.867	0.042	0.540
SIE(color)	0.078	**0.890**	0.097	0.355
SIE(blur)	0.153	0.883	**0.941**	0.189
SIE(trans)	0.213	0.885	0.023	**0.978**
SIE(all)	0.331 ± 0.007	0.899 ± 0.003	0.211 ± 0.005	0.925 ± 0.002
cross_atten_recon	0.893 ± 0.004	0.921 ± 0.006	0.823 ± 0.030	0.875 ± 0.005
rot, inter(angle)	0.9975 ± 0.0005	0.9073 ± 0.0021	0.9310 ± 0.0020	0.9810 ± 0.0010
all, inter(angle)	$\mathbf{0.9983 \pm 0.0005}$	0.9231 ± 0.0005	0.9689 ± 0.0099	0.9801 ± 0.0004
all, inter(color)	0.9891 ± 0.0019	$\mathbf{0.9373 \pm 0.0013}$	0.9700 ± 0.0067	0.9699 ± 0.0022
all, inter(blur)	0.9981 ± 0.0001	0.9154 ± 0.0006	0.9392 ± 0.0106	0.9774 ± 0.0007
all, inter(trans)	0.9975 ± 0.0005	0.9288 ± 0.0012	$\mathbf{0.9747 \pm 0.0017}$	$\mathbf{0.9830 \pm \pm 0.0004}$

- Following SIE [2], we use the coefficient of determination (R^2) as the evaluation metric, measuring how well the model explains outcome variations.
- The head is a multi-layer network that processes pretrained features to output transformation parameters.
- All experiments are fine-tuned on the pretrained model for 50 epochs.

Transfer learning results

Evaluating transfer learning for downstream tasks with Intermediate transformations pretraining

All experiments are based on iBOT [4].



(a) Rotation (b) Translation

- The baseline method is iBOT with ViT-Large pretrained on ImageNet. We measure accuracy improvement using in-between transformed image reconstruction pretraining (rotation and translation), also on ImageNet.
- We evaluate various downstream tasks, including classification (CIFAR-10, FGVC-Aircraft, etc.), dense prediction (semantic segmentation, instance segmentation, object detection, keypoint detection), and homography estimation.
- Classification tasks use a linear head, tuning only the head. Semantic segmentation uses a simple UPerNet [6], while detection and instance segmentation use Mask R-CNN [7]. Dense prediction tasks are fine-tuned for 12 epochs.