Why Do Self-Supervised Models Transfer? On the Impact of Invariance on Downstream Tasks

Linus Ericsson, Henry Gouk and Timothy M. Hospedales

github.com/linusericsson/ssl-invariances

THE UNIVERSITY of EDINBURGH informatics

DATASCIENCE EPSRC CENTRE FOR DOCTORAL TRAINING

UDRC

Royal Academy of Engineering

Samsung Research

Aims

Self-supervised methods have reached a rough agreement on what augmentation optimises popular recognition benchmarks. But different vision tasks likely need different feature (in)variances, and thus different augmentation strategies. In this work, we measure the invariances learned by contrastive methods and their effect on downstream task performances. We pose the following questions:

Q1: Do learned invariances generalise to real-world invariances?
Q2: Is there a trade-off between learning different invariances?
Q3: Do different downstream tasks benefit from different invariances?
Q4: Is there a simple way to achieve high performance across all tasks?

Pre-Training Models

Our main focus is on analysing the properties of self-supervised models pre-trained with different augmentation strategies. We pre-train three models using MoCo-v2 [3] with ResNet50 architectures [6] on ImageNet [4] for 200 epochs.

Default: The default [3, 1, 2, 7, 5] model uses the standard array of data augmentations, which includes crops, horizontal flips, color jitter, grayscale and blur.

Spatial: The Spatial model uses only the spatial subset of default augmentations, including crops and horizontal flips. By learning invariance to these spatial transforms, the model has to put larger focus on colour and texture.

Appearance: The Appearance model uses only the appearance-based augmentations of color jitter, grayscale and blur and will thus have to put larger focus on spatial information.

Apart from differences in augmentation, the pre-training setup is identical for our models. As baselines, we also compare a CNN with **Random** weights, and one pre-trained by Supervised learning on ImageNet.

Evaluation Details

Measuring Invariances: We use two measures of invariance, Mahalanobis distance and cosine similarity. We compute these values between augmented and unaugmented images, averaged over all images considered.

Downstream Tasks: We fit linear models on top of frozen features for a diverse set of downstream tasks. For regression we fit ridge regression and for classification logistic regression.

Results

Synthetic Invariances: The Spatial model is the most invariant to spatial transforms. Likewise, the Appearance model has the strongest invariances to colour and texture. The Default models tends to fall in between these two while the Random model tends to have the highest variance.

Real-World Intrinsic Invariances: We evaluate invariances on datasets that contain real-world transforms, e.g. 3D rotation or lighting changes. The trends between synthetic and real-world invariances are very similar, following the same spatial/appearance split. The interesting outlier is illumination, which goes against the expectation and highlights the importance of understanding the role of data augmentation better.

Real-World Extrinsic Invariances: For a different perspective on invariance to real-world transforms, we regress real-world variables on Causal3DIdent such as pose, object colour and light colour. Results here also confirm that learned invariances translate to some extent to real transformations. The Appearance model obtains better performance on pose prediction tasks, while the Spatial model obtains better performance on colour prediction tasks.

Downstream Tasks: Figure 1 (bottom) shows the linear readout performance on all tasks considered. On the datasets most similar to ImageNet: CIFAR10, Caltech101 and Flowers, the Default or Supervised models achieve the highest classification accuracy, followed by the Spatial and then the Appearance model. On the spatially sensitive tasks the Appearance model outperforms the Spatial model substantively, with the Appearance model performing best overall on 300W. These results show some evidence that the Default (and to a lesser extent Spatial model) model is well suited for object recognition on ImageNet-like datasets, but both are weak in comparison to a model with more spatial sensitivity when solving the pose-related tasks.

Feature Fusion

We explore whether a fusion of specialised features can perform competitively across the board. We explore **Spatial-Appearance (Spa+App)** fusion, as well as three way **Default-Spatial-Appearance (Def+Spa+App)** fusion.

From the results in Fig 1 (3rd & 4th row), we see that the Spa+App model and Def+Spa+App fusion models perform strongly across the board, with the latter being the most consistent performer. While we showed the Default model falls down in this regard, our fused feature provides reliable performance across the board.

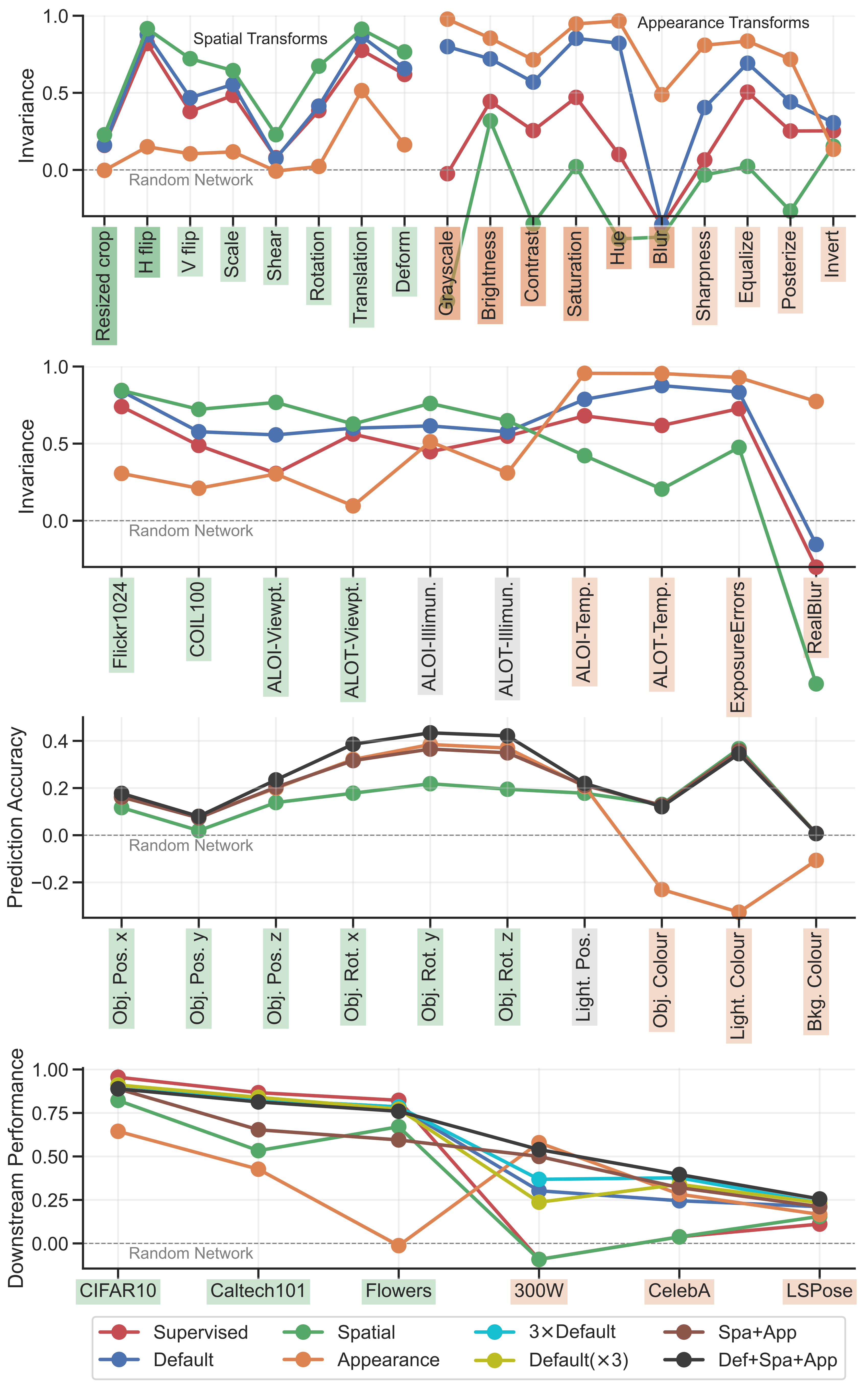

Figure 1: From top to bottom: synthetic invariances, real-world intrinsic invariances, real-world extrinsic invariances, downstream tasks. The y-axes in the top two rows reports cosine similarity in a normalised feature space. For the bottom two rows we report the R^2 regression score or classification accuracy of a linear model on top of frozen features using either ridge or logistic regression, depending on the task.

Conclusions

A1: **Mostly, yes.** E.g., using spatial-style augmentations lead to real-world viewpoint invariance, while appearance-style augmentations lead to increased invariance to lighting colour, exposure and blur. Correspondingly, when predicting pose, having appearance-style invariances help, and vice-versa for predicting colour.

A2: **Yes.** Promoting appearance-style invariances decreases spatial-style ones and vice-versa. We show all existing state-of-the-art learners suffer from this trade-off.

A3: **Yes.** Across a suite of downstream tasks, we see that recognition-style tasks prefer default or spatial-style augmentations, while pose-related tasks benefit from appearance-style augmentations. In particular, default augmentations under-perform in pose-related tasks.

A4: **Yes.** Simple fusion of multiple representations tuned for different (in)variances leads to consistent strong performance across all tasks considered (Fig. 1 third and third fourth row, black line).

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 2 2020.

[2] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *NeurIPS*, 2020.

[3] X. Chen, H. Fan, R. Girshick, and K. He. Improved Baselines with Momentum Contrastive Learning. *arXiv*, 2020.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[5] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In *NeurIPS*, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7] J. Li, P. Zhou, C. Xiong, R. Socher, and S. C. H. Hoi. Prototypical Contrastive Learning of Unsupervised Representations. In *ICLR*, 2021.