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

Linus Ericsson[1]
linus.ericsson@ed.ac.uk
Henry Gouk[1]
henry.gouk@ed.ac.uk
Timothy M. Hospedales[1, 2]
t.hospedales@ed.ac.uk

[1] University of Edinburgh
[2] Samsung AI Research, Cambridge

Abstract

Self-supervised learning is a powerful paradigm for representation learning on unlabelled images. A wealth of effective new methods based on instance matching rely on data-augmentation to drive learning, and these have reached a rough agreement on an augmentation scheme that optimises popular recognition benchmarks. However, there is strong reason to suspect that different tasks in computer vision require features to encode different (in)variances, and therefore likely require different augmentation strategies. In this paper, we measure the invariances learned by contrastive methods and confirm that they do learn invariance to the augmentations used and further show that this invariance largely transfers to related real-world changes in pose and lighting. We show that learned invariances strongly affect downstream task performance and confirm that different downstream tasks benefit from polar opposite (in)variances, leading to performance loss when the standard augmentation strategy is used. Finally, we demonstrate that a simple fusion of representations with complementary invariances ensures wide transferability to all the diverse downstream tasks considered.

1 Introduction

Self-supervised learning has made rapid progress in representation learning, with performance approaching and sometimes surpassing that of supervised pre-training. In computer vision contrastive self-supervised methods driven by data augmentation have been particularly effective [8, 24]. Data augmentation applies synthetic semantics-preserving transformations to images during learning, to increase effective data volume and promote invariance to the augmentation distribution used [13, 14]. By optimising representations so that individual images are similar to their augmented counterparts [9, 22], and possibly also different to alternative distractor images [8, 24, 47], self-supervised algorithms have achieved wide success [16].

In this paradigm the properties and efficacy of the learned representation are largely determined by the augmentation distribution used during self-supervision. To this end a rough consensus has emerged among many state of the art methods as to a good default distribution that leads to strong performance on the downstream benchmarks, especially on the ubiquitous ImageNet object recognition benchmark [14]. For example, image cropping, flipping, colour perturbation

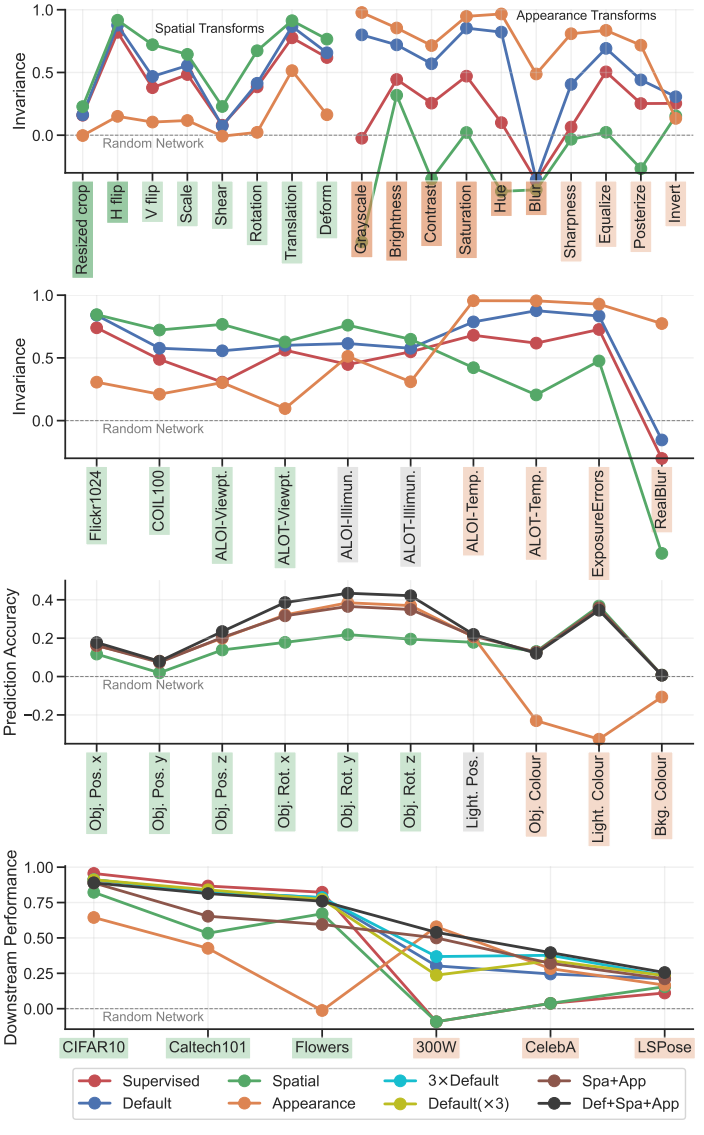

Figure 1: Our Spatial and Appearance models lead to strong spatial and colour/texture invariance respectively, as measured by both synthetic (first row) and real-world (second row) transforms. Simple feature fusion (black) dominates individual pathways, as well as state of the art 'default' augmentation, providing more consistent performance across all downstream tasks (third and fourth row).

and blurring, are widely applied [10, 16, 24]. However, if augmentation leads to invariance to the corresponding transformation, then we should ask: do our self-supervised algorithms provide the right invariances for diverse downstream tasks of interest? For example, while an object categorisation task might benefit from pose invariance, other tasks such as pose estimation may require strong spatial sensitivity. If different tasks require contradictory (in)variances, using a single default data augmentation scheme for all may provide sub-optimal performance for some tasks.

To investigate this issue, we group augmentations into two categories, *spatial* and *appearance*. Using a representative state of the art contrastive learner MoCo-v2+ResNet50 [10], we train models exclusively with spatial-style and appearance-style augmentations and compare them to the model produced by the default augmentation scheme. In particular, we evaluate their resulting invariances to synthetic and real-world transforms, as well as their performance on a suite of diverse real-world downstream tasks.

Based on the experimental design outlined above, we attempt to better understand *why contrastive self-supervised learning works* by answering the following specific questions, among others, with associated results summarised in Figure 1.

Q1 *An increasing amount of work has shown that invariances can be learned by learning augmentations. Do these learned synthetic invariances generalise to real-world invariances?* A1: To some degree, yes. For example, spatial-style augmentations lead to increased invariance to real-world transforms such as viewpoint, while appearance-style augmentations lead to increased invariance to transformations such as lighting colour, exposure and blur. Correspondingly, spatial-style augmentations lead to higher accuracy in estimating object colour, while appearance-style augmentations lead to higher accuracy in estimating object pose. (Fig. 1 second and third row). This has not been measured before.

Q2 *Given that there are multiple types invariances of potential interest to learn. Is there a trade-off between learning different types of invariances?* A2: Yes. Promoting appearance-style invariances decreases spatial-style ones and vice-versa. We also show that all existing state-of-the-art learners suffer from this trade-off.

Q3 *Do different downstream tasks of interest benefit from different invariances?* A3: Yes. Across a suite of downstream tasks, we see that recognition-style tasks prefer a representation trained on default or spatial-style augmentations, while pose-related tasks benefit from appearance-style augmentations. In particular, default augmentations [10] under-perform in pose-related tasks (Fig. 1 fourth row).

Q4 *Given that different tasks prefer polar-opposite augmentations, is there a simple way to achieve high performance across all 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).

2 Related Work

Self-supervision: in computer vision is now too a large topic to review here. Please see [17, 26] for excellent surveys. A key trend is that many highly successful methods rely on matching individual images with augmented versions of themselves, possibly against a background of distractor images. This includes most contrastive methods [2, 24], and some that are not typically considered contrastive [22, 51]. These have been understood [47] as making image features invariant to transformations used for training, while otherwise separating individual images. A key ambition of self-supervision research is for a single pre-trained feature to support diverse downstream tasks, and a common suite of augmentations has emerged to support this [10, 16]. However if data augmentation determines (in)variances, and different tasks require different (in)variances, then a single augmentation distribution may not perform well on all tasks.

Data Augmentation: Data augmentation is the process of transforming input data to increase the diversity of the training set. It has become key to achieving state-of-the-art performance for supervised learning of CNNs in vision [12]. Despite its ubiquity in practice, theoretical understanding of data augmentation is weak. There is some evidence that CNNs can generalise learned translation

invariance [3] to unseen data, but also that they retain information about absolute spatial locations of objects via boundary effects [28].

Data augmentation has become even more vital in practical self-supervision as outlined earlier. However, understanding the role of data augmentation in self-supervision has lagged behind practical engineering lore. Self-supervised contrastive learners with strong augmentation have been shown to learn occlusion-invariant representations, but not to capture viewpoint and category instance invariance [40]. [49] study the theoretical effects of data augmentation on self-supervised contrastive learning. They argue that data augmentation decouples sparse semantic information in the input from dense noisy information and that only the sparse semantic information is relevant to solving the downstream target task. [46] study the effects of data augmentation on invariances and downstream performances using a synthetic task. They show that it is hard to define augmentations to enforce a specific invariance, that augmentations generally have wider invariance effects on groups of factors and that using multiple augmentations in conjunction reliably improves recognition performance. However they focus on object recognition in a synthetic dataset. We take a wider perspective and look at how augmentation impacts a wide variety of real-world transformations, and real-world downstream tasks. A related study to ours is [50], which proposes LooC as a self-supervised method that separates different information into different features, i.e. colour, orientation etc. However, they only evaluate the impact on recognition tasks. A major contribution of ours is to demonstrate how diverse downstream tasks benefit from different learned invariances.

Ventral-Dorsal Visual System: We also briefly highlight an interesting connection between our spatial vs. appearance split and neuroscience. A well established theory about mammalian vision holds that the visual cortex is composed of two functional pathways [20, 30]. The *ventral* stream deals with the "what" of object recognition; and the *dorsal* stream deals with the "where" of spatial and motion information. This decomposition into specialised models has been exploited in applications such as object detection [15] and semantic grasping [25] in robotics. At the intersection of neuroscience and self-supervised learning, [2] showed that a two branch neural network trained with the CPC [37] loss on videos leads to dorsal and ventral-like pathways emerging. Moreover, models of the dorsal stream based entirely on findings from neuroscience and phychophysics (i.e., without use of machine learning) have been shown to accurately estimate motion and depth from videos [11, 39].

We explore self-supervised learning with different data augmentations as a way of achieving similar multi-stream pathways in CNNs for vision. Current methods [7, 24] train representations for invariance to a single set of augmentations that aim to suffice for all tasks. But we show that this current practice is better optimised for the most popular downstream benchmark of object recognition, and poor for pose-related tasks. We will investigate a multi-stream architecture combining representations trained for different invariances, and show it provides more general purpose high performance for diverse downstream tasks.

3 Methods

Our main focus is on analysing the properties of self-supervised models pre-trained with different augmentation strategies. In particular, we choose MoCo-v2 [10] as a representative self-supervised learner that is widely used and near state-of-the-art. MoCo-v2 matches images with their augmented counterparts, while using negative pairs in a contrastive loss to encourage feature dissimilarity between semantic objects, and to avoid features all collapsing to the same vector. We pre-train three models using MoCo-v2 [10] with ResNet50 architectures [23] on ImageNet [12] for 200 epochs.

- **Default:** The default [7, 8, 10, 22, 31] model uses the standard array of data augmentations,

Table 1: Augmentations used during pre-training of our Spatial and Appearance models, along with the standard default augmentations [10]. The color jitter augmentation is a combination of individual jitter in brightness, contrast, saturation and hue.

	Resized crop	Horizontal flip	Color jitter	Grayscale	Blur
Default	✓	✓	✓	✓	✓
Spatial	✓	✓			
Appearance			✓	✓	✓

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.

Table 1 summarises the augmentations used by each model. Apart from these differences, 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.

4 Do Contrastive Methods Learn Invariance to Real-World Transforms?

While several preliminary studies suggest that contrastive methods can learn invariance to synthetic transformations [47], an important question that has not been studied in the literature is whether these learned invariances lead to invariance under real-world transforms, like viewpoint or illumination changes. Does the use of colour augmentations during pre-training lead to features that are invariant to day/night in real images? Does the use of crop/flip augmentation in training lead to pose invariance in real images? While these kinds of questions were intensively studied for classic hand-crafted features [34], they have not been studied for invariances learned by self-supervision. In this section, we investigate whether contrastive methods learn invariance to real-world transforms. We address this question from two perspectives: intrinsically, by measuring the invariance of different representations with respect to different real-world transformations (Section 4.2); and extrinsically, by quantifying how well features trained for different synthetic invariances can be used to predict known real-world transformations (Section 4.3). But first, we provide a more thorough confirmation of the claim that contrastive methods do learn invariances to synthetic transformations.

Measuring invariances: We use two measures of invariance in our experiments, Mahalanobis distance and cosine similarity (full details in Sec C.1 of supplement). We compute these values between augmented and unaugmented images, averaged over all images considered. A further set of measures are reported in Sec C.2 of the supplement with results supporting those in the main paper.

Hypothesis testing: In the following sections we will test several hypotheses based on the cosine similarity invariance measurements. We make use of Hoeffding's inequality, which for a sum of random variables, $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where each X_i is in the range $[0,1]$ with probability one, tells us that

$$P(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-2nt^2}. \tag{1}$$

Table 2: ImageNet pre-trained ResNet50 with MoCo-v2 (200 epochs) evaluated on invariances to transforms on 1000 ImageNet validation images. Top group: Mahalanobis distance where a low value means strong invariance. Bottom group: cosine similarity in a normalised feature space where a value close to 1 means strong invariance. Column colours indicate the type of invariance evaluated and row colours indicate the augmentation expected to lead to high-performing specialised models. The broad agreement between the most invariant features (bold) and expectation (row colours) indicates that training with augmentations does tend to learn the corresponding invariances. Similarity results within {Default, Spatial, Appearance} that are statistically significantly the best are annotated with a •.

		Resized crop	H flip	V flip	Scale	Shear	Rotation	Translation	Deform	Grayscale	Brightness	Contrast	Saturation	Hue	Blur	Sharpness	Equalize	Posterize	Invert
Distance	Random	69.40	34.14	35.35	67.03	69.57	71.65	56.33	65.28	22.81	73.25	59.03	46.63	42.39	49.17	52.59	27.46	27.46	32.88
↓	Supervised	**57.44**	12.33	24.07	40.37	63.93	47.67	22.51	34.25	19.87	40.43	37.05	26.61	35.95	54.92	42.15	17.44	22.74	27.32
	Default	58.72	9.92	21.30	35.75	56.58	43.49	16.45	30.16	7.78	26.07	25.66	12.17	13.72	65.84	31.07	12.81	17.95	**24.71**
	Spatial	59.43	**8.17**	**15.57**	**32.05**	**56.50**	**32.93**	**13.85**	**26.08**	26.58	46.25	61.67	39.34	47.33	63.37	48.83	25.41	38.46	28.95
	Appearance	64.35	27.52	29.05	56.33	71.81	62.49	33.14	52.20	**2.57**	**19.71**	**22.84**	**6.98**	**5.77**	**30.38**	**16.86**	**9.55**	**12.18**	29.24
Similarity	Random	0.03	0.56	0.54	0.16	0.04	0.07	0.40	0.20	0.81	0.17	0.52	0.59	0.60	0.48	0.51	0.68	0.70	0.52
↑	Supervised	0.18	0.92	0.71	0.57	0.11	0.43	0.87	0.69	0.81	0.54	0.64	0.79	0.64	0.29	0.55	0.84	0.77	0.64
	Default	0.19	0.95	0.75	0.63	0.11	0.45	0.92	0.72	0.96	0.77	0.79	0.94	0.93	0.29	0.71	0.90	0.83	**0.67**
	Spatial	**0.25•**	**0.96**	**0.87•**	**0.70•**	**0.26•**	**0.70•**	**0.95•**	**0.81•**	0.65	0.43	0.35	0.60	0.42	0.25	0.50	0.69	0.62	0.59
	Appearance	0.03	0.63	0.59	0.26	0.03	0.09	0.71	0.33	**1.00**	**0.88•**	**0.86•**	**0.98**	**0.99•**	**0.73•**	**0.91•**	**0.95•**	**0.91•**	0.58

Setting the left-hand side equal to δ and rearranging for t yields

$$t \leq \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{2}$$

This fact can be used to test the null hypothesis that the expected value of S_n is zero: set δ to the threshold that will be applied to a p-value, and check whether S_n is greater than the right-hand side of Eq. 2. If it is greater, then one can reject the null hypothesis. By setting S_n equal to the mean difference in representation similarity for two different methods, we can test whether one method is statistically significantly more invariant than the other. Bonferroni correction is applied when we carry out multiple hypothesis tests to perform a three-way comparison.

4.1 Invariance to Synthetic Transforms

Setup: We focus on task-agnostic metrics of invariances. Other extrinsic measures of invariance like identifiability/classification performance under different transformations are inherently biased towards that task. We therefore use invariance metrics that apply to feature vectors directly. We evaluate our Default, Spatial and Appearance methods on 1,000 images from the ImageNet (ILSVRC12) validation set [14] against a wider array of synthetic augmentation transformations than used for training (Tab 1), but still group these into appearance and spatial-style transforms.
Results: The results in Tab. 2 evaluate the invariance of different transformations at test-time (columns) for the different pre-trained models (rows). Using the method described above, we carry out statistical hypothesis tests to determine which of the Default, Appearance, and Spatial models reliably exhibit the most invariance, as measured by the similarity metric. Statistically significant results (at the 95% confidence level) are marked with a •. We make the following observations. For spatial transformations like rotation and translation, the Spatial model is the most invariant, due to its use of such augmentations during pre-training. Likewise, the Appearance model has the strongest invariance to transformations in colour and texture, except for the invert transform. The Default model tends to fall in between the two specialised models suggesting strong invariance to any one transformation is traded off for a reasonable variance across the board. The Random model tends to have the highest variance.

While the spatially-augmented model has very low variance to spatial transforms, it has a high variance to colour and texture. Its sensitivity to these transforms is available for solving tasks that

Table 3: Comparing models in terms of their invariances to real-world transformations. Similarity results within {Default, Spatial, Appearance} that are statistically significantly the best are annotated with a •.

		Flickr1024	COIL100	ALOI	ALOT	ALOI	ALOT	ALOI	ALOT	ExposureErrors	RealBlur
		Stereo	Pose/Scale	Viewpoint	Viewpoint	Illumination	Illumination	Temperature	Temperature	Exposure	Blur
Distance ↓	Random	51.22	39.09	32.36	50.83	33.37	45.99	14.21	41.07	50.14	22.41
	Supervised	27.50	35.40	34.66	43.93	29.72	40.57	10.18	33.03	25.55	25.87
	Default	19.96	24.20	20.29	**39.96**	18.53	37.48	6.60	17.34	17.96	19.15
	Spatial	**19.91**	**23.19**	**15.94**	41.80	**15.73**	**37.28**	11.31	54.28	34.45	32.52
	Appearance	45.37	44.07	38.83	53.62	30.54	41.63	**4.35**	**8.14**	**13.15**	**10.97**
Similarity ↑	Random	0.62	0.42	0.48	0.24	0.58	0.44	0.90	0.73	0.41	0.91
	Supervised	0.90	0.70	0.64	0.67	0.77	0.75	0.97	0.89	0.84	0.89
	Default	**0.94**	0.75	0.77	0.70	0.84	0.76	0.98	0.97	0.90	0.90
	Spatial	**0.94**	**0.84**•	**0.88**•	**0.72**	**0.90**•	**0.80**•	0.94	0.78	0.69	0.82
	Appearance	0.74	0.54	0.64	0.32	0.79	0.61	**1.00**	**0.99**	**0.96**•	**0.98**•

depend on colour or texture. Likewise, the appearance-augmented model is sensitive to spatial information which it could use to solve spatially sensitive tasks. In fact, since the Appearance model is more spatially sensitive than the Default model, it might achieve better performance on such tasks. We investigate this in Sec. 5. Overall the results confirm that invariances are indeed learned by contrastive learning with corresponding augmentations. Furthermore, augmentations do tend to increase invariance to other transforms in the corresponding appearance/spatial family, rather than only the specific subset used for training.

Discussion: We have shown how the use of certain augmentations lead to features that are substantially invariant to those augmentations, as well as others in the same appearance/spatial family. We next address the main question that has not been studied in the literature of whether these learned invariances lead to invariances under real-world transforms, like viewpoint or illumination changes. While these kinds of questions were intensively studied for classic hand-crafted features [34], they have not been studied for invariances learned by self-supervision.

4.2 Real-World Intrinsic Invariance Measurements

Experimental Details: We use the same metrics as in Section 4.1, but instead of using synthetic transformations typically used in data augmentation schemes, we collect a suite of datasets that exhibit known real-world transformations, such as pose changes (Flickr1024, COIL-100. ALOI, ALOT) and colour/appearance changes (ALOI, ALOT, ExposureErrors, RealBlur). Further dataset details can be found in the Sec. D of the supplementary materials. In contrast to the experimental setup in Section 4, we do not have an untransformed reference image. Instead, we consider all pairs of images for a given object/scene/texture within each dataset (or subset of the dataset for ALOI/ALOT), and average our metrics across pairs. For example, in the case of viewing angle variation in COIL-100, a fully pose invariant feature would exhibit full similarity/zero distance across all corresponding image pairs.

Results: The results in Table 3 group the benchmark datasets according to whether they exhibit spatial-like or appearance-like real-world translations. We note that the ALOI/ALOT benchmarks' illumination change condition moves a spotlight in a low ambient light scenario, effectively masking out different parts of the object. As it is unclear whether this corresponds to a appearance or spatial-like transformation, we color these separately. From the results we can see that our models have learned strong real-world invariances in many cases. For example, the spatial model has maximum similarity for all the (green) spatial-like transformations. The appearance model has maximum similarity for all the (orange) appearance-like transformations. We carried out statistical hypothesis tests to determine which of the Default, Spatial, and Appearance models exhibit the most invariance to each type of transformation. These tests were carried out with a

Table 4: Comparing models learned invariances on Causal3DIdent. Regression R^2 fit when predicting parameters from features. Our Appearance model is highly sensitive to spatial-style transforms and our Spatial model highly sensitive to appearance-style transforms. Our fused models exhibit strong predictive capability across the board.

	Obj. Pos. x	Obj. Pos. y	Obj. Pos. z	Obj. Rot. x	Obj. Rot. y	Obj. Rot. z	Spot. Pos.	Obj. Colour	Spot. Colour	Bkg. Colour
Spatial	0.89	0.91	0.84	0.66	0.67	0.62	0.92	**0.93**	**0.91**	**1.00**
Appearance	0.94	**0.97**	0.91	0.80	0.84	0.79	0.95	0.56	0.22	0.88
Spa+App	0.94	**0.97**	0.91	0.80	0.82	0.77	0.95	0.92	0.90	**1.00**
Def+Spa+App	**0.95**	**0.97**	**0.94**	**0.87**	**0.89**	**0.85**	**0.96**	0.92	0.89	**1.00**

confidence level of 95%, and statistically significant maxima are marked with a •. That the spatial model trained with crops and flips achieves stronger illumination invariance than one trained with colour augmentations on the ALOI/ALOT datasets is interesting, and highlights the importance of understanding the role of data augmentation better in representation learning.

Summary: In summary, we asked *Q1: Whether invariances learned using data augmentation generalise to real-world transforms?* Grouping by appearance and spatial family invariances and real-world transforms, the answer is yes. To visualise this qualitatively, we report in Figure E.1 of our supplementary materials the rank correlation (over our five models) between all pairs of synthetic and real-world invariances considered. From the results we can see a clear block structure that strong invariance to appearance-like synthetic transforms correlates with the appearance-like real-world transforms, and vice-versa. More formally, the statistical significance tests in Table 3 show that five of eight comparisons show statistically significant impact of training augmentation with real-world transformation invariance.

For *Q2: Is there a trade-off between learning different types of invariances?* A2: Yes. We have found that both for synthetic and real-world transforms, increasing appearance-style invariances decreases spatial-style ones and vice-versa. Models using the default set of augmentations suffer from this trade-off (we show that other state-of-the-art learners suffer similarly in Fig. F.1 of the supplement) Next, we look into whether a greater invariance to one family of transforms induces sensitivity in the other.

4.3 Real-World Extrinsic Invariance Measurements

To provide a different perspective on invariance to real-world transforms, we evaluate our features on the Causal3DIdent benchmark. In particular, we follow [46] in regressing real-world variables such as pose, object colour, light colour, etc. from our features. A feature with complete spatial invariance would fail to predict pose, while one with colour invariance would fail to predict colour, etc.

Setup: We use kernel ridge regression with an RBF kernel, sample 20,480 training points and 40,960 test points and standardise images and targets. As the feature dimensionality of our models is much greater than of those in [46], we expand the hyperparameter search space for α and γ to $[10^{-4},10^{-3},10^{-2},10^{-1},10^{0}]$ and $[10^{-5},10^{-4},10^{-3},10^{-2},10^{-1},10^{0},10^{1},10^{2}]$, respectively.

Results: From the results in Table 4 this evaluation paradigm also confirms 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.

Table 5: Downstream performances of our models. We report mean and standard deviation of 5-fold cross-validation on all data for each task. The differing performances in the tasks showcases how the Spatial and Appearance models capture important but *different* properties necessary for wide transfer. Random refers to a randomly initialised feature extractor and '+' refers to feature concatenation. On the left datasets we report the classification accuracy and on the right the R^2 regression metric. Row colours indicate whether appearance or spatial turned out better for the given task.

	CIFAR10	Caltech101	Flowers	300W	CelebA	LSPose	Avg.
Random	0.55 ± 0.004	0.25 ± 0.008	0.21 ± 0.009	0.24 ± 0.024	0.47 ± 0.002	0.10 ± 0.007	0.30 ± 0.009
Supervised	$\mathbf{0.98 \pm 0.001}$	$\mathbf{0.90 \pm 0.005}$	$\mathbf{0.86 \pm 0.007}$	0.17 ± 0.028	0.49 ± 0.002	0.20 ± 0.015	0.60 ± 0.010
Default	0.96 ± 0.002	0.87 ± 0.006	0.83 ± 0.004	0.47 ± 0.014	0.60 ± 0.002	$\mathbf{0.29 \pm 0.025}$	$\mathbf{0.67 \pm 0.009}$
Spatial	0.92 ± 0.003	0.65 ± 0.008	0.74 ± 0.010	0.17 ± 0.030	0.49 ± 0.001	0.24 ± 0.020	0.54 ± 0.013
Appearance	0.84 ± 0.003	0.57 ± 0.007	0.20 ± 0.009	$\mathbf{0.68 \pm 0.018}$	$\mathbf{0.62 \pm 0.003}$	0.25 ± 0.021	0.53 ± 0.010
3×Default	$\mathbf{0.96 \pm 0.004}$	0.87 ± 0.004	$\mathbf{0.83 \pm 0.007}$	0.52 ± 0.012	0.67 ± 0.001	0.31 ± 0.016	0.69 ± 0.007
Default(×3)	$\mathbf{0.96 \pm 0.002}$	$\mathbf{0.88 \pm 0.003}$	0.82 ± 0.009	0.42 ± 0.030	0.65 ± 0.005	0.31 ± 0.028	0.67 ± 0.013
Spa+App	0.95 ± 0.002	0.74 ± 0.009	0.68 ± 0.005	0.62 ± 0.021	0.64 ± 0.002	0.29 ± 0.007	0.65 ± 0.008
Def+Spa+App	0.95 ± 0.003	0.86 ± 0.009	0.81 ± 0.006	$\mathbf{0.65 \pm 0.020}$	$\mathbf{0.68 \pm 0.002}$	$\mathbf{0.33 \pm 0.010}$	$\mathbf{0.71 \pm 0.008}$

5 Do Downstream Tasks Prefer Different Invariances?

In the previous sections we have showed how contrastive training under data augmentation learns invariance to synthetic and real-world transformations. It also confirmed the colour/texture sensitivity of the Spatial model and the spatial sensitivity of the Appearance model. The Default model was found to always fall in between the two more specialised learners, with weaker invariance than one alternative but stronger than the other.

In terms of real-world benchmarks, self-supervised methods are widely evaluated on ImageNet recognition, with the literature having a lesser focus and lack of consistency in evaluation of other non-recognition tasks. Since the default augmentations are largely chosen to optimise recognition benchmarks, there is a chance that it may be overfit to these tasks and perform less well on others. We therefore investigate how learned invariances affect a more diverse suite of real downstream tasks of interest, hypothesising that different features may be preferred, depending on the (in)variance needs of each downstream task.

Experimental Details: Our suite of downstream tasks consists of object recognition on standard benchmarks **CIFAR10** [29], **Caltech101** [18] and **Flowers** [36]; as well as a set of spatially sensitive tasks including facial landmark detection on **300W** [42] and **CelebA** [33], and pose estimation on **Leeds Sports Pose** [27]. We freeze the backbones and extract features from just after the average pooling layer of the ResNet50 architectures. We fit a ridge or logistic regression model on these features, depending on the task in question. To tune the $\ell 2$ regularisation value we perform 5-fold cross-validation over a grid of 45 logarithmically spaced values between 10^{-6} to 10^{5}, following [7, 16]. We report the mean and standard deviation for the hyperparameter choice with highest mean. The performance is reported as accuracies (between 0 and 1) for classification tasks and R^2 values for regression tasks. For comparison we also evaluate random and supervised backbones.

Results: Table 5 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. Overall this supports the hypothesis that different (in)variances

are required for best performance on different types of tasks. To answer *Q3: Do different downstream tasks of interest benefit from different invariances?* A3: Yes. On our classification tasks, representations trained on default or spatial-style augmentations dominate. Pose-related tasks benefit from appearance-style augmentations, where default augmentations under-perform.

Improving Performance Through Feature Fusion: Our previous analyses show that different real-world tasks prefer different invariances. The Default model tries to satisfy them all by using a mix of augmentations to obtain a moderate amount of invariance to all transformations (Sec 4), but appearance/spatial specialised features can be better for particular tasks (Table 5, top). We therefore explore whether a fusion of specialised features can perform competitively across the board. In particular we explore Spatial-Appearance (Spa+App) fusion, as well as three way Default-Spatial-Appearance (Def+Spa+App) fusion.

Experimental Details: The evaluation follows the setup described above, but as our fused representations have higher dimensionality, we shift the $\ell 2$ search space for Spa+App to 10^{-5} to 10^{6} and Def+Spa+App 10^{-4} to 10^{7}. Finally, to compare the concatenated features of Def+Spa+App, we evaluate a second Default model with a $3\times$ wider architecture – ResNet50($\times 3$) – which was trained with MoCo-v2 for 200 epochs on ImageNet like our other models and uses the same hyperparameter search space as Def+Spa+App. A final baseline consisting of three fused separately trained Default models forms $3\times$Default.

Results: From the results in Table 5 (bottom) and Table 4 (bottom), we can see that the Spa+App model and Def+Spa+App fusion models perform strongly across the board. While the $3\times$Default and Default($\times 3$) models are unsurprisingly best for recognition tasks, this is only by a small margin; while the 3-way Def+Spa+App fusion is dramatically better for 300W, and the most consistent performer across the board. To answer *Q4: Is there a simple way to achieve high performance across all tasks?* A4: Yes. We fuse multiple representations tuned for different (in)variances for consistently strong performance across all tasks considered.

This result is noteworthy, as a goal of self-supervised learning is to provide a single feature that provides excellent performance for diverse downstream tasks. While we showed the default model falls down in this regard, our fused feature provides reliable performance across the board. We therefore recommend it to practitioners who want a single feature with which to perform diverse tasks.

6 Discussion

We have performed the first thorough evaluation of self-supervised learning in terms of augmentations used for training, and resulting downstream invariance and task impact. In particular we showed that: (1) CNNs trained contrastively do learn invariances corresponding to the augmentations used and specialising CNNs to particular appearance/spatial augmentations can lead to greater corresponding invariances (Table 2). Furthermore, learning invariances to synthetic transforms does provide a degree of invariance to corresponding real-world transforms (Table 3, Fig E.1). (2) Different real-world downstream tasks do prefer features providing different invariances (Table 5, Fig. F.1), and invariance-specialised features can sometimes outperform the standard default augmentation, e.g., for spatially sensitive tasks. (3) Fusing features tuned for different types of invariances provides a consistently high performing strategy (Table 5). This outperforms the default model on pose related tasks, suggesting that it was over-tuned for recognition. Our feature ensemble strategy is promising for providing high performance *general purpose* real-world features. Based on these results we encourage the SSL community to evaluate on more diverse downstream task types.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/S000631/1; and the MOD University Defence Research Collaboration (UDRC) in Signal Processing. This project was supported by the Royal Academy of Engineering under the Research Fellowship programme.

References

[1] Mahmoud Afifi, Konstantinos G. Derpanis, Björn Ommer, and Michael S. Brown. Learning Multi-Scale Photo Exposure Correction. In *CVPR*, 2021.

[2] Shahab Bakhtiari, Patrick Mineault, Tim Lillicrap, Christopher C. Pack, and Blake A. Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In *NeurIPS*, 2021.

[3] Valerio Biscione and Jeffrey Bowers. Learning Translation Invariance in CNNs. In *2nd Workshop on Shared Visual Representations in Human and Machine Intelligence, NeurIPS*, 2020.

[4] Gertjan J Burghouts and Jan-Mark Geusebroek. Material-specific adaptation of color invariant features. *Pattern Recognition Letters*, 2009.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *NeurIPS*, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 2020.

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *NeurIPS*, 2020.

[9] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *CVPR*, 2021.

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv*, 2020.

[11] Michael J Cree, John A Perrone, Gehan Anthonys, Aden C Garnett, and Henry Gouk. Estimating heading direction from monocular video sequences using biologically-based sensors. In *IVCNZ*, 2016.

[12] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning Augmentation Policies from Data. In *CVPR*, 2019.

[13] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020.

[14] 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.

[15] Mohammad K. Ebrahimpour, Jiayun Li, Yen-Yun Yu, Jackson L. Reese, Azadeh Moghtaderi, Ming-Hsuan Yang, and David C. Noelle. Ventral-Dorsal Neural Networks: Object Detection via Selective Attention. In *WACV*, 2019.

[16] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How Well Do Self-Supervised Models Transfer? In *CVPR*, 2021.

[17] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances and challenges. *IEEE Signal Processing Magazine*, 2022.

[18] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004.

[19] Jan-Mark Geusebroek, Gertjan J. Burghouts, and Arnold W.M. Smeulders. The Amsterdam Library of Object Images. *International Journal of Computer Vision*, 2005.

[20] Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 1992.

[21] Ian J Goodfellow, Quoc V Le, Andrew M Saxe, Honglak Lee, and Andrew Y Ng. Measuring Invariances in Deep Networks. In *NeurIPS*, 2009.

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In *NeurIPS*, 2020.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, 2019.

[25] Eric Jang, Sudheendra Vijayanarasimhan, Peter Pastor, Julian Ibarz, and Sergey Levine. End-to-End Learning of Semantic Grasping. In *CoRL*, 2017.

[26] Longlong Jing and Yingli Tian. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[27] Sam Johnson and Mark Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*, 2010.

[28] Osman Semih Kayhan and Jan C. van Gemert. On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location. In *CVPR*, 2020.

[29] Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. *arXiv*, 2009.

[30] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[31] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical Contrastive Learning of Unsupervised Representations. In *ICLR*, 2021.

[32] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[34] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[35] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL100). Technical report, 1996.

[36] Maria Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[37] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. *NeurIPS*, 2017.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 2019.

[39] John A. Perrone, Michael J. Cree, and Mohammad Hedayati. Using the properties of primate motion sensitive neurons to extract camera motion and depth from brief 2-d monocular image sequences. In *CAIP*, 2019.

[40] Senthil Purushwalkam and Abhinav Gupta. Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases. In *NeurIPS*, 2020.

[41] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-World Blur Dataset for Learning and Benchmarking Deblurring Algorithms. In *ECCV*, 2020.

[42] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 Faces In-The-Wild Challenge. *Image and Vision Computing*, 2016.

[43] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In *GCPR*, 2014.

[44] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 2019.

[45] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking Representation Learning for Natural World Image Collections. In *CVPR*, 2021.

[46] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *NeurIPS*, 2021.

[47] Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *ICML*, 2020.

[48] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a Ladder: A New Understanding of Contrastive Learning. In *ICLR*, 2022.

[49] Zixin Wen and Yuanzhi Li. Toward Understanding the Feature Learning Process of Self-supervised Contrastive Learning. In *ICML*, 2021.

[50] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What Should Not Be Contrastive in Contrastive Learning. In *ICLR*, 2021.

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *ICML*, 2021.