Supplementary

Bram Wallace and Bharath Hariharan

Cornell University
bw462@cornell.edu

1 Experimental Setup

1.1 Architectures

ResNet26 See Figure 3 and Section 3.2 of [1] for the original description.

Autoencoder Generator The architecture is laid out in Table S1.

Layer	Output Shape
Input	256
ConvTranspose2d-1	[-1, 512, 4, 4]
BatchNorm2d-2	[-1, 512, 4, 4]
ReLU-3	[-1, 512, 4, 4]
ConvTranspose2d-4	[-1, 256, 8, 8]
BatchNorm2d-5	[-1, 256, 8, 8]
ReLU-6	[-1, 256, 8, 8]
ConvTranspose2d-7	[-1, 128, 16, 16]
BatchNorm2d-8	[-1, 128, 16, 16]
ReLU-9	[-1, 128, 16, 16]
ConvTranspose2d-10	[-1, 64, 32, 32]
BatchNorm2d-11	[-1, 64, 32, 32]
ReLU-12	[-1, 64, 32, 32]
ConvTranspose2d-13	[-1, 3, 64, 64]
Tanh-14	[-1, 3, 64, 64]

Table S1. Generator architecture. First convolution has stride of 1 and no padding, all subequent convolutions have stride of 2 with padding 1. All kernels have size 4.

2 Training & Evaluation

All networks are trained using stochastic gradient descent for 120 epochs with an initial learning rate of 0.1 decayed by a factor of 10 at 80 and 100 epochs, with momentum of 0.9. One addition to our training process was that of "Earlier

Stopping" for the Rotation and Jigsaw pretext tasks. We found that even with traditional early stopping, validation accuracy could oscillate as the pretext overfit to the training data (especially in the Scenes & Textures or Biological cases), potentially resulting in a poor model as the final result. We stabilized this behavior by halting training when the training accuracy improves to 98%, effect on accuracy shown in Table S2.

Table S2. Comparison of test accuracies with early stopping vs without. Rotation in particular was stabilized and improved by this method. Jigsaw was stabilized, but sometimes hampered. For Jigsaw with less permutations than the 2000 reported the net effect was more positive. The only qualitative difference in results was Jigsaw matching Instance Discrimination on the Internet domains instead of being outperformed. Both methods still fell far behind Rotation.

	Jigsaw Early	Jigsaw Regular	Rotation Early	Rotation Regular
aircraft	8	9	9	11
cifar100	19	24	42	37
cub	9	9	12	14
daimlerpedcls	67	80	87	87
dtd	15	14	15	14
gtsrb	68	67	82	79
isic	57	59	60	62
merced	57	53	70	58
omniglot	18	24	46	54
scenes	33	33	42	40
svhn	50	53	80	78
ucf101	25	22	42	45
vgg-flowers	22	19	23	22
bach	47	46	41	36
protein atlas	21	21	22	25
kather	79	78	57	61

2.1 Dataset Splits

We use provided dataset splits when available, taking our validation data from training data when a train-validation split is not predetermined.[1] If no split was given, we generally used a 60-20-20 split within each class. *Full train-validation-test splits will be released along with our code and models.*

2.2 Data Augmentation, Weight Decay, and Other Regularization

A sensitive topic in any deep learning comparison is that of data augmentation or other forms of regularization, which can substantially alter performance. In

[1] Despite using overlapping domains with the VDC, we are forced to use different splits in some cases due to the Visual Decathlon challenge not releasing the corresponding test labels.

this work we are determined to give as fair of an apples-to-apples comparison as possible, and as such we apply minimal data augmentation and do not employ weight decay or other regularization methods to the main paper results. In experiments with weight decay of $5e-4$, we found that Autoencoding and Instance Discrimination improved by 3 and 4% respectively. These trials were performed with weight decay instead of Earlier Stopping, and both Rotation and Jigsaw actually performed *worse* with this traditional regularization (by 8 and 0.5%) respectively.

The data augmentation used consists solely of resizing, random crops, and horizontal flips. Note that horizontal flips are not typically used on the symbolic domains, but are considered standard everywhere else. We elected to go with the logical choice for 13 out of 16 of our domains, and employ horizontal flips in all of our main experiments. We present results without flipping below.

2.3 Effect of Horizontal Flipping on Symbolic

As seen in Table S3, taking away horizontal flipping generally does not have major effects *except* for improving Rotation-Omniglot substantially and hurting Jigsaw-SVHN significantly. The former we attribute to the learning load of Rotation being used, while the latter we posit is due to the lack of horizontal flips allowing Jigsaw to use simpler cues for classification.

Table S3. Each tuple is normal accuracy (with horizontal flips, as in paper) and accuracy without flips. In general we see performance changes of only a few percentage points, qualitative comparisons largely hold. The biggest differences are Rotation's improvement on Omniglot and Jigsaw's worsening on SVHN.

	Autoencoding	Jigsaw	ID	Rotation	Supervised
GTSRB	(57,58)	(66, 67)	(43, 39)	(82, 78)	(93, 93)
SVHN	(31, 33)	(55, 26)	(37, 34)	(80, 81)	(95, 95)
Omniglot	(18,19)	(26, 27)	(45, 47)	(46, 53)	(79, 80)

3 Implicit Dimensionality

We observe that the largest variations in explained variance between pretexts occur in the first dimension (Table S4), and investigate its use as a predictor in downstream performance. Correlations are shown in Figure S1. We do observe a moderate correlation between the explained variance in the first component and downstream normalized accuracy for Instance Discrimination. While weak, this trend holds for PCA performed on both the training and validation images. More significantly, we note the distinct separation formed around 0.5 on the x-axis and perform a t-test to determine that there is a moderately significant

difference in downstream accuracies across this interval ($p = 0.052$). Thus implicit dimensionality is mildly predicitive of downstream performance for Instance Discrimination.

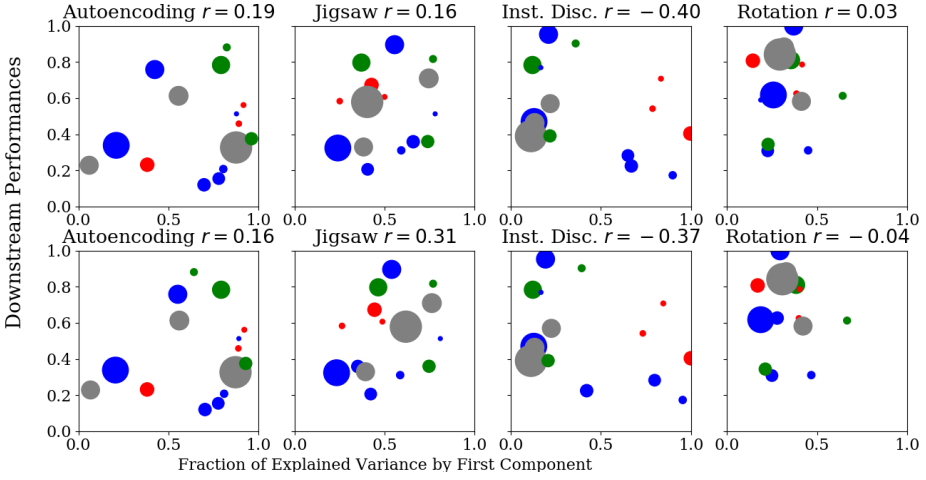

Fig. S1. Downstream normalized classification accuracy vs. the fraction of variance explained by the first component. Top row is PCA on the entire training feature set, the bottom on validation. The only moderately significant trends are those of Instance Discrimination, but we note that the trend holds with comparable strength for both sets.

4 Correlations of Pretexts with Downstream Accuracy

Correlations for each task are shown in Figures S2, S4, S3, S5. X-axis is accuracy for Rotation/Jigsaw, loss of Autoencoding and ID.

References

1. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: Advances in Neural Information Processing Systems (2017)

Table S4. Fraction variance explained by the first n values.

	Autoencoder		Jigsaw		Inst. Disc.		Rotation	
n	256	4096	256	4096	256	4096	256	4096
1	0.67	0.47	0.53	0.29	0.43	0.35	0.33	0.15
2	0.75	0.54	0.69	0.39	0.51	0.41	0.51	0.23
3	0.79	0.58	0.78	0.46	0.56	0.45	0.60	0.29
4	0.82	0.61	0.83	0.50	0.61	0.48	0.67	0.33
5	0.84	0.64	0.86	0.53	0.66	0.51	0.71	0.36
10	0.89	0.72	0.92	0.63	0.79	0.61	0.82	0.46
15	0.92	0.77	0.94	0.69	0.86	0.66	0.87	0.53
20	0.94	0.81	0.95	0.72	0.90	0.70	0.90	0.57
30	0.96	0.85	0.97	0.76	0.94	0.76	0.93	0.63
40	0.97	0.88	0.97	0.79	0.96	0.79	0.95	0.67
50	0.98	0.90	0.98	0.81	0.97	0.82	0.96	0.71
60	0.99	0.92	0.98	0.83	0.98	0.84	0.96	0.73
70	0.99	0.92	0.99	0.84	0.98	0.85	0.97	0.75
80	0.99	0.93	0.99	0.85	0.99	0.86	0.97	0.77
90	0.99	0.94	0.99	0.86	0.99	0.88	0.98	0.78
100	0.99	0.94	0.99	0.87	0.99	0.88	0.98	0.80
110	1.00	0.94	0.99	0.87	0.99	0.89	0.98	0.81
120	1.00	0.95	0.99	0.88	0.99	0.90	0.99	0.82
130	1.00	0.95	0.99	0.88	1.00	0.91	0.99	0.82
140	1.00	0.95	1.00	0.89	1.00	0.91	0.99	0.83
150	1.00	0.95	1.00	0.89	1.00	0.92	0.99	0.84

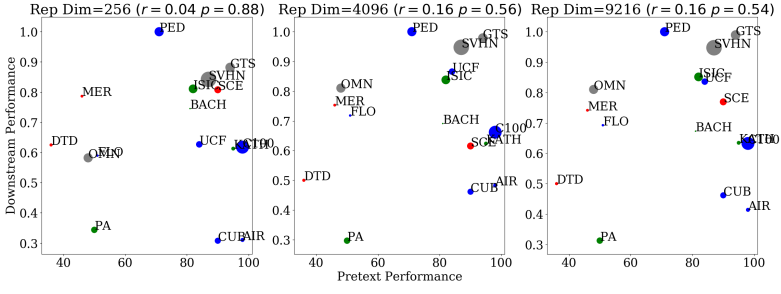

Fig. S2. Downstream normalized classification accuracy vs. performance on pretext task for Rotation.

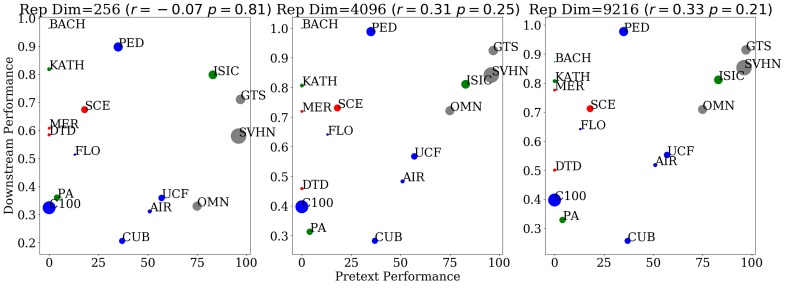

Fig. S3. Downstream normalized classification accuracy vs. performance on pretext task for Jigsaw.

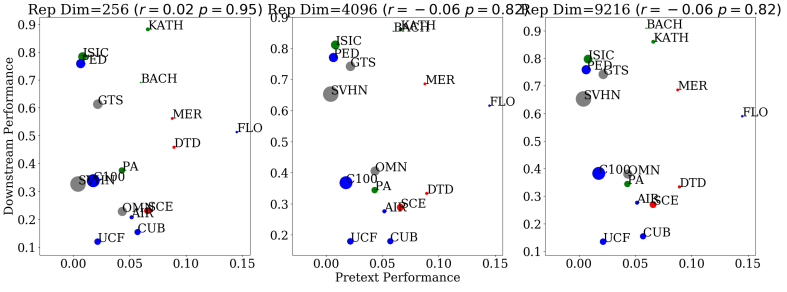

Fig. S4. Downstream normalized classification accuracy vs. performance on pretext task for Autoencoding.

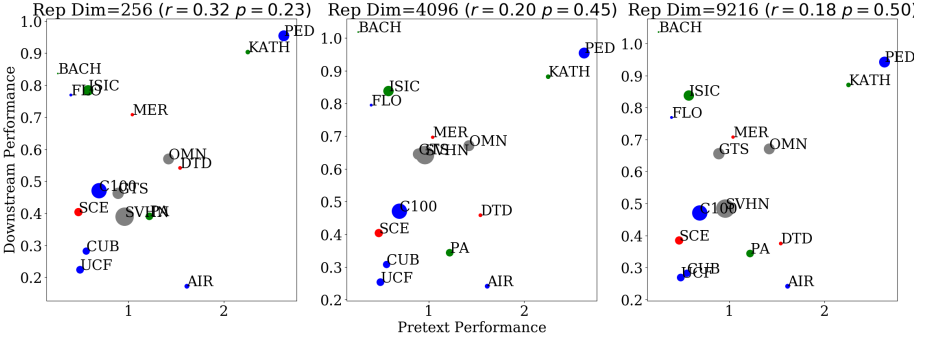

Fig. S5. Downstream normalized classification accuracy vs. performance on pretext task for Instance Discrimination.