HIDDEN-LAYER SELF-DISTILLATION YIELDS DRIFT-RESILIENT VISUAL REPRESENTATIONS

Scott C. Lowe[1,*] **Anthony Fuller**[1,2], **Sageev Oore**[1,3], **Graham W. Taylor**[1,4], **Evan Shelhamer**[1,5]
[1]Vector Institute, [2]Carleton Uni., [3]Dalhousie Uni., [5]Uni. of Guelph, [4]Uni. of British Columbia

ABSTRACT

The choice of pretraining strategy has a direct impact on how well visual representations withstand distribution shift at deployment. We study **Bootleg**, a recent self-supervised method that predicts continuous latent representations from *multiple hidden layers* of an EMA teacher network, spanning early stimulus-driven features to late semantic features. This multi-scale objective forces representations to encode both fine-grained spatial detail and high-level semantics. We evaluate Bootleg against MAE, CrossMAE, data2vec 2.0, and I-JEPA across three ViT scales (S, B, L) on distribution-shift benchmarks. We find that Bootleg pretraining yields best or second best robustness across all model sizes. We further show that Bootleg representations respond well to test-time adaptation with SAR, yielding the largest accuracy gains under corruption shift. These results suggest that grounding SSL targets across the network hierarchy is a promising strategy for drift-resilient representation learning.

1 INTRODUCTION

Deploying vision models in the real world exposes them to distribution shifts—changes in illumination, weather, sensor noise, artistic style, and object context that differ from training data. A model's resilience to such drift depends not only on the downstream adaptation strategy but also on the *pretraining objective* that shaped its representations. Self-supervised learning (SSL) methods such as masked autoencoder (MAE) (He et al., 2022) and I-JEPA (Assran et al., 2023) are widely adopted for their data-agnostic masking paradigms, yet their robustness properties under distribution shift remain underexplored.

We study Bootleg (Lowe et al., 2026), a recent SSL method that tasks the model with predicting representations from multiple *hidden layers* of an exponential moving average (EMA) teacher network, rather than reconstructing pixels (as in MAE) or predicting only final-layer embeddings (as in I-JEPA). By targeting outputs spread across the teacher's depth—from early, stimulus-driven layers to late, semantic layers—Bootleg forces the encoder to build representations that are simultaneously grounded in both low-level structure and rich in high-level semantics. This compression of hierarchical structure into the embeddings may provide a natural buffer against distribution shift: early-layer targets anchor the representation to input statistics, while late-layer targets enforce semantic invariance.

Our contributions are: **(1)** A systematic robustness evaluation of masked SSL methods across three vision transformer scales (ViT-S, -B, -L) on 19 unseen datasets (VTAB) and five distribution-shift benchmarks. **(2)** Evidence that fine-tuned Bootleg models achieve the best-in-class robustness among masked-SSL methods. **(3)** Demonstration that Bootleg representations respond most favourably to test-time adaptation with SAR (Niu et al., 2023), yielding the largest accuracy gains for probes of frozen pretrained models.

2 BACKGROUND

Bootleg (Lowe et al., 2026) extends the I-JEPA framework (Assran et al., 2023) by expanding the self-distillation prediction targets from the teacher's final layer to multiple hidden layers. **Masking.**

[*]Correspondence: scott.lowe@vectorinstitute.ai

Table 1: **VTAB benchmark results.** Category-average top-1 accuracy (%) across Natural (7 tasks), Specialized (4 tasks), and Structured (8 tasks) categories. We report category-averages for Patch, CLS, and X-Blk frozen probes of encoders pretrained on IN-1k with either masking-based SSL, or cross-entropy supervision (Sup.). **Best** and second-best highlighted in each category.

		Natural			Specialized			Structured			Overall		
Arch	Method	Patch	CLS	X-Blk	Patch	CLS	X-Blk	Patch	CLS	X-Blk	Patch	CLS	X-Blk
ViT-S	Sup. (AugReg)	60.5	58.9	60.4	**78.9**	**78.6**	**78.7**	33.1	**31.6**	**41.7**	52.8	**51.5**	**56.4**
	Sup. (DeiT III)	**61.3**	**60.7**	**61.3**	77.6	76.4	76.6	**33.8**	30.0	39.7	**53.1**	51.1	55.4
	MAE	45.7	46.2	57.6	76.6	77.3	78.2	39.1	38.9	53.2	49.5	49.7	60.1
	CrossMAE	47.9	46.4	59.7	77.4	76.0	78.8	**41.2**	**40.2**	**57.5**	51.3	50.1	62.8
	data2vec 2.0	38.8	37.4	48.4	74.6	75.6	74.1	40.1	38.4	54.7	46.9	45.8	56.5
	I-JEPA	40.2	N/A	45.8	74.3	N/A	73.8	37.7	N/A	43.9	46.3	N/A	50.9
	Bootleg	**59.8**	**59.8**	**66.2**	**79.0**	**78.1**	**80.8**	40.4	38.9	52.5	**55.6**	**54.9**	**63.5**
ViT-B	Sup. (torch)	61.6	65.0	68.3	**80.6**	**79.2**	**82.1**	**37.6**	33.4	47.5	**55.5**	54.7	62.5
	Sup. (AugReg)	62.5	60.8	65.8	79.9	78.6	79.8	34.8	32.2	40.6	54.5	52.5	58.1
	Sup. (DeiT III)	**67.4**	**65.1**	**69.6**	80.0	77.6	79.8	32.1	**34.6**	**47.9**	55.2	**54.9**	**62.6**
	MAE	54.3	53.4	64.4	76.7	76.8	79.2	42.2	42.2	58.7	53.9	53.6	65.1
	CrossMAE	56.4	55.2	65.8	77.6	74.3	81.0	**45.7**	44.7	**61.8**	56.4	54.8	**67.3**
	data2vec 2.0	52.6	43.9	54.6	76.6	71.6	72.4	43.7	**44.9**	53.8	53.9	50.2	58.0
	I-JEPA	50.4	N/A	55.1	77.2	N/A	78.5	39.8	N/A	49.8	51.6	N/A	57.8
	Bootleg	**60.7**	**61.5**	**68.0**	**82.1**	**80.5**	**83.7**	39.8	38.0	57.8	**56.4**	**55.6**	67.0
ViT-L	Sup. (torch)	67.0	64.3	71.4	**81.5**	**80.5**	**82.2**	35.3	33.2	46.0	56.7	54.6	63.0
	Sup. (DeiT III)	**69.6**	**68.4**	**72.2**	81.2	79.4	81.1	**39.6**	**35.1**	**48.6**	**59.4**	**56.7**	**64.1**
	MAE	56.2	60.1	65.7	79.4	80.2	81.6	45.4	44.7	62.1	56.6	**57.9**	67.5
	CrossMAE	59.1	57.6	68.4	78.8	**80.5**	81.4	46.0	46.5	54.1	**57.7**	57.8	65.1
	data2vec 2.0	51.9	47.9	55.5	77.8	74.0	80.1	**46.3**	**46.5**	**62.5**	55.0	52.8	63.6
	I-JEPA	54.0	N/A	58.0	78.4	N/A	79.1	38.0	N/A	46.3	52.4	N/A	57.5
	Bootleg	**60.7**	**64.9**	**70.9**	**81.9**	79.7	**84.2**	41.5	39.7	58.9	57.1	57.4	**68.7**

Block-structured masks (four rectangular regions), as in I-JEPA, are applied to the image. **Student-Encoder.** Processes visible (unmasked) patches, augmented with one CLS token and four register tokens (Darcet et al., 2024). **Targets** A teacher-encoder (using exponential moving average weights of the student-encoder) processes the full (unmasked) image. Embeddings are collected from every 4th block (e.g., blocks 1, 4, 8, 12 for ViT-B), independently z-scored, and concatenated into a single target vector per masked patch. **Predictor.** A ViT that uses student-encoder outputs to map mask tokens to the multi-layer target, with four predictor-only register tokens for global processing. **Loss.** Mean squared error between predictor's mask token outputs and multi-layer targets.

I-JEPA training (Assran et al., 2023) is the same as Bootleg, except targets only come from the final layer of the teacher network; also the predictor is shallower, warmup shorter, weight decay higher.

MAE training methodology (He et al., 2022) is the same as Bootleg, except the training targets are the image pixels, the masking strategy is uniform random, and warmup is shorter.

CrossMAE training (Fu et al., 2025) is the same as MAE, except self-attention in the predictor is replaced with cross-attn and it sees a learnt weighted average of the output of every encoder block.

data2vec 2.0 training (Baevski et al., 2023) uses the average of the last 10 blocks as the training target, runs 16 masking repetitions of each image on each step, with a smaller, convolutional predictor.

3 Experiments

We evaluate ViT models pretrained on ImageNet-1k (IN-1k) with 224×224 images and 16×16 patches using SSL methods. We additionally show fully-supervised (cross-entropy) baselines AugReg (Steiner et al., 2022), DeiT-III (Touvron et al., 2022), and torchvision-builtin model as upper bounds.

Versatility of pretrained embeddings. We used frozen probes to evaluate pretrained models on the Visual Task Adaptation Benchmark (VTAB) (Zhai et al., 2019), using the protocol of 1000

Table 2: **Robustness of fine-tuned models under distribution shift.** Models pretrained on IN-1k with SSL, fine-tuned on IN-1k, evaluated on shifted test sets. For each shifted dataset, we show (black) the top-1 accuracy (%) on the dataset and (red) delta compared with IN-1k. For context, we additionally show the performance of fully-supervised models (Sup.) trained on IN-1k from scratch. **Best** and second-best highlighted in each category.

Arch	Method	Ep.	IN-1k	IN-V2		IN-R		IN-Sketch		IN-A		IN-C	
ViT-S	Sup. (AugReg)	300	74.9	62.4	−12.5	33.2	−41.7	17.1	−57.8	8.8	−66.1	25.5	−49.4
	Sup. (DeiT III)	400	**81.3**	**70.8**	**−10.6**	**46.5**	**−34.8**	**35.0**	**−46.3**	**27.2**	**−54.2**	**42.7**	**−38.6**
	MAE	800	78.2	66.5	−11.6	42.8	−35.4	28.5	−49.7	13.6	−64.6	32.7	−45.5
	CrossMAE	800	79.8	68.6	−11.2	43.8	−36.0	30.3	−49.5	18.4	−61.4	34.6	−45.2
	data2vec 2.0	200	79.9	69.2	**−10.7**	45.8	−34.1	31.3	−48.6	19.1	−60.7	36.9	−42.9
	I-JEPA	600	79.9	69.2	−10.7	44.5	−35.4	31.4	−48.5	21.4	−58.5	38.8	−41.1
	Bootleg	600	**80.8**	**70.0**	−10.8	**47.7**	**−33.1**	**33.9**	**−46.9**	**23.9**	**−56.9**	**39.9**	**−40.9**
ViT-B	Sup. (torch)	300	81.0	69.6	−11.5	44.1	−37.0	29.4	−51.7	20.8	−60.2	39.5	−41.5
	Sup. (AugReg)	300	76.8	64.1	−12.6	37.4	−39.4	22.6	−54.2	11.3	−65.5	31.4	−45.3
	Sup. (DeiT III)	400	**83.7**	**73.6**	**−10.1**	**53.6**	**−30.1**	**40.1**	**−43.6**	**39.9**	**−43.8**	**50.7**	**−32.9**
	MAE	1600	82.7	72.2	−10.5	45.6	−37.1	33.2	−49.5	32.6	−50.1	42.4	−40.3
	CrossMAE	800	82.9	72.5	−10.4	48.1	−34.8	35.0	−47.9	35.9	−47.0	44.3	−38.6
	data2vec 2.0	200	83.3	73.7	**− 9.7**	50.2	−33.1	37.7	**−45.6**	37.2	−46.2	45.6	−37.7
	I-JEPA	600	82.7	71.3	−11.4	47.8	−34.8	34.6	−48.0	32.1	−50.5	**46.4**	**−36.2**
	Bootleg	600	**83.9**	**74.0**	− 9.9	**51.2**	**−32.7**	**37.9**	−46.1	**39.6**	**−44.3**	46.4	−37.6
ViT-L	Sup. (torch)	600	79.7	67.5	−12.2	40.8	−38.9	26.8	−52.9	17.4	−62.3	40.8	−38.9
	Sup. (DeiT III)	400	**84.6**	**74.8**	**− 9.8**	**56.8**	**−27.8**	**43.3**	**−41.2**	**49.0**	**−35.5**	**56.1**	**−28.5**
	MAE	1600	84.7	74.8	− 9.8	55.4	−29.2	41.4	−43.3	50.6	−34.0	51.0	−33.6
	CrossMAE	800	84.3	74.7	− 9.6	53.3	−31.0	39.7	−44.6	48.0	−36.4	50.7	−33.6
	data2vec 2.0	200	**85.6**	**76.9**	**− 8.7**	**59.1**	**−26.5**	**47.5**	**−38.1**	**57.8**	**−27.8**	**58.5**	**−27.2**
	I-JEPA	600	82.0	71.1	−10.9	48.6	−33.4	35.7	−46.3	28.4	−53.6	46.2	−35.8
	Bootleg	600	85.4	75.8	− 9.6	57.3	−28.1	43.6	−41.8	53.7	−31.7	54.4	−31.0

training samples per task (see Appx. B). As shown in Table 1, Bootleg achieves the highest overall accuracy across all probe types at ViT-S and is best or second-best at ViT-B and ViT-L. The gains are most pronounced on Natural tasks (most similar to the training domain) where Bootleg leads other SSL models by a wide margin, and on Specialized tasks (out-of-domain images such as satellite and medical images) where Bootleg consistently beats both SSL *and* supervised pretrained models. On Structured tasks, which require spatial reasoning and counting with far-OOD synthetic images, CrossMAE and data2vec 2.0 are more competitive.

Fine-tuned model robustness. When models are fine-tuned on IN-1k and evaluated on distribution-shifted test sets (Table 2), Bootleg consistently achieves the strongest or second strongest performance among SSL methods across all model sizes, both in terms of maximizing accuracy on the distribution-shifted data, and in minimizing reduction in performance compared to in-domain data.

Test-time adaptation with SAR. To evaluate how well representations respond to drift at test time, we apply SAR (Niu et al., 2023)—a backpropagation-based entropy minimization method that adapts encoder LayerNorm parameters using a sharpness-aware optimizer—to frozen linear probes fit on IN-1k. As shown in Table 3, Bootleg consistently achieves the highest post-adaptation accuracy among SSL methods, with the largest Δ, indicating its representations are most amenable to test-time correction. These results suggest that multi-layer targets produce feature spaces with smoother loss landscapes that SAR can more effectively exploit.

We also observe the performance of data2vec 2.0 decreases with SAR. This is could potentially be due to an architectural difference in these pretrained models rather than indicative of the SSL method: the architecture is post-norm instead of pre-norm, and SAR operates by adapting the norm parameters. But we note that for fine-tuned models (Appx. D.2), data2vec 2.0 responds favourably to SAR.

4 Discussion and Future Work

Why does multi-layer distillation help robustness? We hypothesize two complementary mechanisms. First, *multi-scale grounding*: by targeting early layers (which are stimulus-driven and spatially

Table 3: **Test-time adaptation with SAR of frozen linear probes.** IN-C shows the mean accuracy over 15 corruptions. **Best** and second-best methods highlighted.

		IN-R			IN-C		
Arch	Method	Base	SAR	Δ	Base	SAR	Δ
ViT-S/16	Supervised (AugReg)	33.2	34.4	+1.1	25.5	44.8	**+19.3**
	Supervised (DeiT III)	**46.5**	**51.1**	**+4.5**	**42.7**	**47.8**	+5.1
	MAE	16.6	17.2	+0.6	6.1	4.1	−2.0
	CrossMAE	20.4	21.0	+0.6	8.3	5.8	−2.5
	data2vec 2.0	12.1	12.1	+0.0	5.0	3.7	−1.3
	I-JEPA	14.2	15.0	+0.9	7.1	7.9	+0.8
	Bootleg	**26.9**	**28.6**	**+1.7**	**15.4**	**17.3**	**+1.9**
ViT-B/16	Supervised (torch)	44.1	39.6	−4.4	39.5	49.1	+9.6
	Supervised (AugReg)	37.4	39.8	**+2.4**	31.4	53.4	**+22.0**
	Supervised (DeiT III)	**53.6**	**55.9**	+2.3	**50.7**	**61.0**	+10.3
	MAE	26.4	28.2	+1.8	11.3	12.0	+0.7
	CrossMAE	27.9	29.6	+1.7	13.2	13.7	+0.5
	data2vec 2.0	24.5	23.5	−1.0	13.5	11.2	−2.3
	I-JEPA	19.7	21.3	+1.6	11.5	13.7	+2.2
	Bootleg	**31.6**	**33.9**	**+2.3**	**19.4**	**23.5**	**+4.1**
ViT-L/16	Supervised (torch)	40.8	41.1	+0.3	40.8	45.4	+4.5
	Supervised (DeiT III)	**56.8**	**58.6**	**+1.8**	**56.1**	**63.7**	**+7.6**
	MAE	31.3	33.4	+2.1	20.7	23.4	+2.7
	CrossMAE	31.8	33.9	+2.1	19.6	24.1	+4.5
	data2vec 2.0	31.2	25.7	−5.5	**34.8**	14.6	−20.2
	I-JEPA	20.6	22.1	+1.5	13.9	18.5	+4.7
	Bootleg	**34.3**	**37.0**	**+2.6**	25.3	**30.1**	**+4.8**

detailed) alongside deep layers (which are abstract and semantic), the encoder receives training signal is anchored to input statistics at multiple levels. Under distribution shift, early-layer targets help resist low-level corruption (noise, blur), while late-layer targets maintain semantic invariance across domain changes (style, context). Second, the *information bottleneck* (Tishby et al., 1999): with $|L|$ target layers, width D, $|M|$ masked patches, and $|N|$ visible patches, the predictor must reconstruct $|L| \times |M| \times D$ elements from a bottleneck of $|N| \times D_{\text{pred}}$ (with $|N| < |M|$), forcing the encoder to build richer, more structured representations that are less likely to overfit to distribution-specific artifacts.

Conclusions. We have demonstrated that Bootleg representations provide the best backbone for generalizing to unseen natural and specialized VTAB images, and it yields fine-tuned models the more robust in the face of distribution shift on IN-1k. The most robust ViT-L model was data2vec 2.0, which also uses multi-layer teacher representations to construct its targets, though by averaging final layers instead of concatenating representations from across the visual hierarchy like Bootleg. These findings suggest multi-scale pretraining produces representations which are more stable. Bootleg also adapts more effectively under test-time adaptation with SAR, suggesting that multi-scale pretraining produces feature spaces amenable to test-time correction.

Future work. The core contribution of Bootleg is to use multiple representations of different levels of abstraction as training targets. The implementation can be transferred readily to any JEPA-family model across modalities (Bardes et al., 2024; Assran et al., 2025; Fei et al., 2024; Tuncay et al., 2025), opening up the possibility of improved robustness for models across many domains and scales.

Layer agreement—measuring consistency between early- and late-layer predictions—is a promising direction for unsupervised drift sensing. Combining Bootleg with continual learning could enable models that maintain robustness as the data distribution evolves over time. During Bootleg pretraining, the predictor learns to map between representations at different abstraction levels. At deployment, divergence between early- and late-layer prediction errors could serve as an unsupervised drift indicator without requiring labelled data.

ACKNOWLEDGEMENTS

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. doi:10.1109/CVPR52729.2023.01499.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. doi:10.48550/arxiv.2506.09985.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1416–1429. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/baevski23a.html.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. doi:10.48550/arxiv.2404.08471. Featured Certification.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-JEPA: Joint-embedding predictive architecture can listen. *arXiv preprint arXiv:2311.15830*, 2024. doi:10.48550/arxiv.2311.15830.

Letian Fu, Long Lian, Renhao Wang, Baifeng Shi, XuDong Wang, Adam Yala, Trevor Darrell, Alexei A. Efros, and Ken Goldberg. Rethinking patch dependence for masked autoencoders. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=JT2KMuo2BV.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi:10.1109/CVPR52688.2022.01553.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, Los Alamitos, CA, USA, October 2021a. IEEE Computer Society. doi:10.1109/ICCV48922.2021.00823.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, June 2021b.

Scott C. Lowe, Anthony Fuller, Sageev Oore, Evan Shelhamer, and Graham W. Taylor. Self-distillation of hidden layers for self-supervised representation learning. *arXiv preprint arXiv:2603.15553*, 2026. doi:10.48550/arxiv.2603.15553.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=4nPswr1KcP.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999. doi:10.48550/arXiv.physics/0004057.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 516–533, Cham, 2022. Springer Nature Switzerland. doi:10.1007/978-3-031-20053-3_30. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136840509.pdf.

Ludovic Tuncay, Etienne Labbé, Emmanouil Benetos, and Thomas Pellegrini. Audio-JEPA: Joint-embedding predictive architecture for audio representation learning. *arXiv preprint arXiv:2507.02915*, 2025. doi:10.48550/arxiv.2507.02915.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3eefceb8087e964f89c2d59e8a249915-Paper.pdf.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. doi:10.48550/arxiv.1910.04867.

APPENDICES

A EVALUATION METHODOLOGY

A.1 FROZEN PROBES

After SSL pretraining, probes of the frozen encoders were trained as follows. (1) **Patch**: linear probe on average patch embeddings (discarding CLS and register tokens), with batch norm; (2) **CLS**: linear probe on the CLS token embedding (discarding patch and register tokens), with batch norm; (3) **X-Blk**: attentive probe following V-JEPA (Bardes et al., 2024; Assran et al., 2025)—a cross-attention block with a single learnable query, an MLP sub-block, and a linear head, using the same width and attention heads as the pretrained ViT.

All probes are trained with cross-entropy for 20 epochs, batch size 1024, sweeping 25 learning rate and weight decay configurations. We use random crop $[0.3, 1.0]$ with basic augmentations only (horizontal flip, colour jitter). As in He et al. (2022), the final encoder norm is discarded. See Lowe et al. (2026, Appendix E.1) for further details.

A.2 FINE-TUNING

Our methodology for full-fine tuning on IN-1k follows that recommended in the GitHub repository of MAE (He et al., 2022) and used by CrossMAE (Fu et al., 2025): heavy augmentation (RandAugment, Mixup, Cutmix, RandomErasure, drop path), average patch embeddings with layer norm and linear head, label-smoothed cross-entropy; learning rate 5×10^{-4} (ViT-S,-B), 3×10^{-4} (ViT-L data2vec 2.0 only*), or 1×10^{-3} (ViT-L); LR layer decay of 0.65 (ViT-S,-B) or 0.75 (ViT-L); trained for 50 (ViT-L) or 100 (ViT-S,-B) epochs.

*Note: we found data2vec 2.0 ViT-L collapsed during fine-tuning with the recommended learning rate from He et al. (2022) and so reduced it by a factor of three. This adds some unfairness to the results on fine-tuned models in its favour as other models are only evaluated at a single LR value (which is presumably already the optimal value for MAE, as it is their protocol) instead of two LR values.

B VTAB CLASSIFICATION

We evaluate on the Visual Task Adaptation Benchmark (VTAB) (Zhai et al., 2019), which comprises 19 diverse visual classification tasks grouped into three categories. **Natural**: 7 standard recognition tasks set in the natural world such as CIFAR-100, Caltech-101, and Flowers-102. **Specialized**: 4 domain-specific tasks including medical scans, microscopy, and satellite imagery. **Structured**: 8 tasks requiring geometric or spatial reasoning, such as object counting and orientation estimation, using synthetic images.

B.1 METHOD

Following the VTAB methodology, we train our probes for 50 epochs on a 1000 sample subset of the training set only, without data augmentation. We use the same probe types as for IN-1k (Patch, CLS, X-Blk) described in Appx. A.1, evaluated on the full test set for each of the 19 datasets.

The training hyperparameters (learning rate and weight decay) are selected by training a model on a fixed subset of 800 training samples, and validated on the remaining 200 samples. The best performing probes is then trained again from scratch on the full set of 1000 training samples. The evaluation hyperparameters considered are shown in Table 4. Due to the diversity of the tasks, we needed a wide range of learning rate and weight decay hyperparameters to attain optimal results across the models and dataset pairs; consequently, we sweep an 11×11 grid of learning rate and weight decay values for each head, using a factor of 3 spacing: $\eta_{\text{max}} \in \{3\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}, \cdots, 3.0\}$ $\lambda \in \{0, 1\times10^{-4}, 3\times10^{-4}, \cdots, 3.0\}$.

Table 4: **Hyperparameter configuration for VTAB frozen probes.** We sweep an 11×11 grid of learning rates and weight decay values for each probe type.

	Hyperparameter	Value
Data	Input size	224×224
	Interpolation	Bicubic
	Augmentations	None
Training	Loss	Cross-entropy
	Optimizer	AdamW
	LR schedule	Cosine decay
	Warmup	5 epochs
	LR initial	2×10^{-4}
	LR max	$3\times10^{-5} \leq \eta_{\text{max}} \leq 3.0$
	LR final	0.0
	WD	$0 \leq \lambda \leq 3.0$
	Batch size	64
	Training samples	1000
	Epochs	50

B.2 RESULTS

Table 1 presents category-average results across Natural, Specialized, and Structured task groups. Bootleg achieves the highest overall accuracy across all three probe types at ViT-S and is best or second-best at ViT-B and ViT-L. The gains are most pronounced on Natural tasks, where Bootleg leads by a wide margin (*e.g.*, 59.8% vs. 47.9% Patch at ViT-S), and on Specialized tasks, where Bootleg consistently achieves top scores. On Structured tasks, which require spatial reasoning and counting, CrossMAE and data2vec 2.0 are more competitive. At ViT-L, Bootleg's X-Blk probe achieves the highest overall accuracy (68.7%), with MAE and CrossMAE close behind on individual categories.

A break-down of results on individual datasets is shown for a linear Patch probe in Table 5 and for an attentive X-Blk probe in Table 6.

Table 5: VTAB benchmark results: per-dataset top-1 accuracy (%) using a Patch (linear) frozen probe. C100 = CIFAR-100, Cal. = Caltech-101, Flr. = Flowers-102, Cam. = Camelyon17, Eur. = EuroSAT, Res. = Resisc45, Ret. = Retinopathy, Cl = Clevr, dS = dSprites, sN = SmallNORB, DM = DMLab, Kit. = KITTI. Sup. (torch) = torchvision builtin model.

	Natural							Specialized				Structured							
Method	C100	Cal.	DTD	Flr.	Pets	Sun	SVHN	Cam.	Eur.	Res.	Ret.	Cl_C	Cl_D	dS_L	dS_O	sN_A	sN_E	DM	Kit.
ViT-S																			
Sup. (AugReg)	38.5	**84.1**	**61.6**	**79.4**	85.0	32.9	**42.1**	78.5	**89.4**	**73.5**	**74.1**	**39.6**	**40.5**	**27.4**	**22.8**	**13.6**	24.4	**33.9**	62.3
Sup. (DeiT III)	**40.8**	82.9	60.1	78.5	**88.9**	**37.3**	40.2	**80.9**	88.3	67.3	73.9	38.2	39.6	25.8	20.6	11.5	**36.0**	32.7	**65.8**
MAE	24.1	74.1	54.0	65.3	35.6	18.4	48.6	76.3	90.8	65.7	73.8	42.4	46.4	44.0	31.0	18.5	**39.4**	33.6	57.7
CrossMAE	26.6	76.4	54.4	68.3	45.0	15.8	48.8	**78.9**	89.6	67.2	73.8	47.8	50.6	45.4	**34.5**	**19.6**	34.3	33.5	**63.9**
data2vec 2.0	20.3	59.9	47.1	60.4	24.4	15.2	44.3	73.5	88.1	63.4	73.5	44.3	**51.9**	**54.4**	18.9	16.5	39.4	33.2	62.0
I-JEPA	15.9	68.2	44.2	58.4	46.1	16.0	32.3	71.6	87.4	64.8	73.6	43.7	43.9	53.4	20.8	15.7	32.3	34.4	57.0
Bootleg	**35.8**	**87.4**	**62.9**	**80.2**	**70.9**	**32.0**	**49.1**	73.3	**92.3**	**75.9**	**74.4**	**49.5**	48.0	40.9	31.0	18.9	35.1	**36.2**	63.4
ViT-B																			
Sup. (torch)	41.6	84.2	65.3	83.9	76.0	31.1	49.0	77.4	**92.5**	**78.4**	74.0	42.2	**45.7**	**42.6**	26.3	17.8	**33.3**	34.5	**58.6**
Sup. (AugReg)	41.8	85.4	61.9	81.7	85.5	37.5	43.9	77.8	92.1	75.2	74.5	43.6	43.1	34.6	24.9	17.1	26.8	32.5	56.1
Sup. (DeiT III)	**48.7**	**87.5**	**66.3**	**87.1**	**89.4**	**41.5**	**51.3**	**78.2**	89.8	77.5	**74.6**	**43.7**	39.6	33.3	**33.0**	**18.3**	31.4	**35.6**	22.1
MAE	27.1	84.5	60.5	73.8	62.7	25.0	46.4	79.5	90.2	71.2	65.8	50.5	48.5	43.8	34.2	20.3	36.8	34.5	69.5
CrossMAE	32.6	83.9	59.2	75.1	62.1	24.0	57.7	77.1	90.8	68.8	73.6	49.5	52.6	48.8	**46.2**	**24.5**	**39.5**	34.3	**70.6**
data2vec 2.0	32.7	80.3	56.1	66.3	44.6	24.1	**64.5**	75.2	90.2	67.0	73.9	**50.7**	**56.5**	**50.0**	29.0	23.9	37.6	32.6	69.1
I-JEPA	20.9	78.1	49.5	69.5	**78.9**	21.3	34.8	78.0	88.7	67.9	74.1	47.3	45.4	46.6	26.5	20.5	35.5	32.8	64.0
Bootleg	**35.1**	**87.8**	**64.0**	**82.7**	75.2	**34.9**	45.2	**82.4**	**92.5**	**78.5**	**74.9**	49.0	48.1	40.6	33.3	15.8	39.0	**36.0**	56.8
ViT-L																			
Sup. (torch)	49.0	87.6	67.3	87.0	88.0	38.9	51.1	**79.1**	**92.8**	80.4	73.7	40.8	41.7	26.9	25.1	**17.6**	30.8	34.2	65.1
Sup. (DeiT III)	**52.3**	**89.5**	**69.9**	**88.0**	**89.8**	**41.4**	**56.6**	77.0	92.3	**80.7**	**74.6**	**44.1**	**43.4**	**34.7**	**34.1**	17.6	**34.3**	**37.2**	**71.4**
MAE	28.5	84.8	62.7	75.5	62.3	26.4	53.5	80.7	91.1	72.9	73.0	52.6	49.9	48.1	43.4	23.8	39.5	34.5	**71.6**
CrossMAE	36.4	85.5	**63.0**	76.5	66.6	27.5	**58.0**	**81.6**	91.8	68.3	73.4	52.8	50.3	**51.9**	**49.4**	24.0	39.8	33.2	66.4
data2vec 2.0	37.5	82.4	54.0	67.3	41.8	23.0	57.8	80.2	90.3	69.7	71.0	**57.3**	**51.3**	50.2	38.4	**24.7**	**44.4**	36.5	67.8
I-JEPA	29.4	78.9	54.1	69.6	**78.8**	28.1	39.5	76.6	91.9	71.6	73.5	43.6	46.0	44.8	25.3	20.0	34.2	33.1	57.0
Bootleg	**39.2**	**88.0**	58.8	**85.1**	69.5	**34.5**	49.9	78.8	**92.8**	**80.9**	**75.2**	49.0	50.1	43.2	34.8	20.8	38.6	**37.2**	58.4

Table 6: VTAB benchmark results: per-dataset top-1 accuracy (%) using an X-Blk (attentive) frozen probe. C100 = CIFAR-100, Cal. = Caltech-101, Flr. = Flowers-102, Cam. = Camelyon17, Eur. = EuroSAT, Res. = Resisc45, Ret. = Retinopathy, Cl = Clevr, dS = dSprites, sN = SmallNORB, DM = DMLab, Kit. = KITTI. Sup. (torch) = torchvision builtin model.

	Natural							Specialized				Structured							
Method	C100	Cal.	DTD	Flr.	Pets	Sun	SVHN	Cam.	Eur.	Res.	Ret.	Cl_C	Cl_D	dS_L	dS_O	sN_A	sN_E	DM	Kit.
ViT-S																			
Sup. (AugReg)	36.6	79.4	**59.5**	**79.5**	85.5	31.8	**50.4**	78.6	**89.5**	**73.7**	**72.8**	**55.2**	**52.1**	**51.5**	**30.2**	**15.8**	26.0	**35.0**	67.5
Sup. (DeiT III)	**39.1**	**83.5**	58.7	79.1	**89.5**	**33.2**	46.1	**80.1**	88.4	69.6	68.3	53.9	46.6	34.8	23.9	11.5	**46.0**	33.0	**68.1**
MAE	27.8	80.2	54.0	75.6	68.5	16.7	**80.0**	75.4	92.6	71.3	**73.7**	72.6	58.9	73.5	45.0	24.7	36.7	41.0	73.1
CrossMAE	29.8	84.2	54.7	77.6	76.6	18.6	76.6	79.3	92.0	70.6	73.0	76.5	58.1	**83.9**	**50.9**	**32.8**	36.4	43.2	**78.1**
data2vec 2.0	19.1	64.2	46.0	69.0	55.0	12.9	72.5	73.0	88.1	69.0	66.2	69.2	**62.1**	81.4	43.9	25.8	**43.9**	41.4	70.2
I-JEPA	19.3	73.2	43.0	61.0	64.8	13.9	45.1	70.5	86.3	66.0	72.3	53.6	57.0	68.1	22.1	21.6	29.0	34.9	65.0
Bootleg	**36.0**	**87.8**	**61.1**	**86.5**	**84.9**	**31.2**	75.7	**81.6**	**93.0**	**80.2**	68.3	**77.9**	61.1	73.5	47.7	25.9	40.6	**43.5**	49.9
ViT-B																			
Sup. (torch)	46.3	**88.5**	67.0	**87.9**	89.8	34.5	63.9	**83.8**	**93.2**	**79.6**	**72.0**	64.3	53.8	**63.9**	36.4	14.2	34.1	**40.1**	73.4
Sup. (AugReg)	42.6	86.7	62.3	83.3	87.2	37.7	60.5	78.9	91.9	76.9	71.5	61.7	**56.5**	6.4	**41.1**	**21.4**	29.4	39.2	69.3
Sup. (DeiT III)	**48.8**	87.8	**67.1**	87.4	**90.8**	**40.6**	**64.6**	80.6	90.0	77.1	71.3	**67.0**	49.0	63.0	38.0	19.6	**35.1**	36.3	**75.1**
MAE	30.1	86.2	60.0	82.6	84.8	24.9	82.5	79.4	91.2	75.4	70.7	79.4	62.0	84.1	48.0	30.7	41.7	45.6	77.8
CrossMAE	35.8	**89.1**	59.2	83.4	84.2	25.1	**84.0**	83.9	93.2	76.4	70.6	**85.8**	61.1	**86.1**	**55.9**	**34.6**	**44.8**	**47.6**	78.3
data2vec 2.0	27.4	85.1	47.7	61.7	59.9	16.5	83.9	75.5	88.1	61.0	65.0	74.2	**62.7**	76.2	38.2	25.8	42.5	40.8	70.2
I-JEPA	23.1	76.9	50.5	76.1	77.8	22.1	59.0	75.9	90.7	75.5	72.1	71.0	57.1	68.3	30.3	26.1	35.4	38.1	72.0
Bootleg	**38.4**	87.9	**64.5**	**89.7**	**89.5**	**35.8**	70.3	**84.5**	**93.9**	**83.5**	**72.8**	80.7	60.1	69.8	49.7	29.3	44.7	45.5	**82.7**
ViT-L																			
Sup. (torch)	49.2	88.3	67.8	87.8	89.2	39.4	**78.1**	82.0	**92.9**	**81.8**	71.9	62.9	**54.8**	49.6	**46.3**	5.4	35.4	**43.5**	70.6
Sup. (DeiT III)	**53.1**	**89.7**	**70.9**	**88.4**	**90.8**	**42.8**	69.9	**84.0**	92.6	75.6	**72.4**	**67.1**	49.5	**60.4**	26.2	**23.1**	**44.3**	37.1	**80.9**
MAE	31.6	89.1	61.7	83.8	88.2	24.1	81.6	83.3	92.5	78.0	72.5	91.2	62.9	**87.3**	51.2	32.7	46.2	47.0	78.5
CrossMAE	40.0	90.0	63.2	85.3	85.3	27.8	**87.4**	82.6	93.0	77.2	72.9	**92.9**	60.8	6.4	**58.1**	**36.3**	45.9	**48.7**	83.5
data2vec 2.0	42.3	85.4	53.8	80.9	20.4	24.2	81.6	81.8	92.1	77.0	69.4	89.4	**64.0**	86.3	44.7	28.6	**56.2**	46.1	**84.5**
I-JEPA	30.4	80.0	54.4	73.3	82.8	28.4	56.6	77.8	91.2	73.5	74.0	69.0	57.5	62.1	13.2	21.2	32.9	41.3	73.4
Bootleg	**45.2**	**90.6**	**67.6**	**92.6**	**89.6**	**35.6**	75.1	**83.4**	**94.5**	**84.7**	**74.2**	84.8	62.1	71.8	52.8	28.7	46.2	47.7	77.5

C CLASSIFICATION ROBUSTNESS UNDER DOMAIN SHIFT

C.1 METHODOLOGY

We first take the SSL-pretrained model and fine-tune it for IN-1k classification as per Appx. A.2. Supervised pretrained models are used as-is without modification.

Using the classifier head trained on IN-1k, we deploy it without modification on several datasets. These datasets contain the same semantic classes as ImageNet-1k, but have some degree of domain shift to the imagery:

- IN-V2: Close replication of IN-1k distribution (Recht et al., 2019) (matched-frequency-format).
- IN-R (Rendition): Artworks of various types, representing IN-1k classes (Hendrycks et al., 2021a).
- IN-Sk (Sketch): Sketches of IN-1k classes (Wang et al., 2019).
- IN-A (Adversarial): Naturally occurring images of IN-1k classes which confuse models (Hendrycks et al., 2021b).
- IN-C (Corruption): IN-1k validation images distorted with 15 different types of corruption (Hendrycks & Dietterich, 2019).

For each dataset, we measure the mean accuracy of the classifier. For IN-C, we evaluate on each corruption type at severity 5.

C.2 ROBUSTNESS OF FROZEN PROBES

To accompany Table 2, we show the robustness of linear probes and attentive probes atop SSL-pretrained models. We take the SSL-pretrained encoder and, while keeping it frozen, train linear and attentive probes atop of it to predict IN-1k classes (see Appx. A.1). We then evaluate the linear/attentive IN-1k classifier head on domain-shifted variants as described above. Supervised

pretrained models are still used as-is without modification, and are exactly the same method and results as above.

For the linear probes, Bootleg is the best performing model across all distribution-shifted datasets (Table 7, black values). However, this is not because it was most *resistant* to a reduction in performance, per-say. Bootleg has a much higher linear-probe performance on IN-1k than the other SSL methods, so when its performance is reduced by a comparatively large amount (Table 7, red values) it simply retains the top-spot on the shifted datasets.

Table 7: **Frozen linear probe robustness.** Performance of patch-avg linear probe trained on IN-1k and evaluated on various robustness benchmarks, compared with cross-entropy supervised baselines. **Best** and second-best highlighted in each category.

Arch	Method	Ep.	IN-1k	IN-V2		IN-R		IN-Sketch		IN-A		IN-C	
ViT-S	Sup. (AugReg)	300	74.9	62.4	−12.5	33.2	−41.7	17.1	−57.8	8.8	−66.1	25.5	−49.4
	Sup. (DeiT III)	400	**81.3**	**70.8**	**−10.6**	**46.5**	**−34.8**	**35.0**	**−46.3**	**27.2**	**−54.2**	**42.7**	**−38.6**
	MAE	800	47.0	35.3	−11.7	16.6	−30.3	9.2	−37.8	1.7	−45.2	6.1	−40.9
	CrossMAE	800	51.8	39.6	−12.2	20.4	−31.4	12.0	−39.8	1.7	−50.1	8.3	−43.5
	data2vec 2.0	200	39.9	29.4	**−10.5**	12.1	**−27.8**	5.3	**−34.6**	1.6	**−38.3**	5.0	**−34.9**
	I-JEPA	600	52.4	40.0	−12.4	14.2	−38.2	5.5	−47.0	2.4	−50.0	7.1	−45.3
	Bootleg	600	**69.8**	**56.4**	−13.3	**26.9**	−42.9	**17.7**	−52.1	**4.7**	−65.0	**15.4**	−54.4
ViT-B	Sup. (torch)	300	81.0	69.6	−11.5	44.1	−37.0	29.4	−51.7	20.8	−60.2	39.5	−41.5
	Sup. (AugReg)	300	76.8	64.1	−12.6	37.4	−39.4	22.6	−54.2	11.3	−65.5	31.4	−45.3
	Sup. (DeiT III)	400	**83.7**	**73.6**	**−10.1**	**53.6**	**−30.1**	**40.1**	**−43.6**	**39.9**	**−43.8**	**50.7**	**−32.9**
	MAE	1600	66.1	52.7	−13.3	26.4	−39.7	16.3	−49.7	3.8	−62.3	11.3	−54.7
	CrossMAE	800	65.5	52.7	−12.8	27.9	**−37.6**	17.9	−47.6	3.6	−61.9	13.2	−52.3
	data2vec 2.0	200	62.2	49.3	−12.8	24.5	−37.7	18.2	**−44.0**	3.2	**−59.0**	13.5	**−48.6**
	I-JEPA	600	67.0	53.6	−13.3	19.7	−47.3	11.0	−56.0	4.0	−62.9	11.5	−55.5
	Bootleg	600	**75.5**	**62.9**	**−12.6**	**31.6**	−43.9	**22.1**	−53.4	**8.9**	−66.6	**19.4**	−56.1
ViT-L	Sup. (torch)	600	79.7	67.5	−12.2	40.8	−38.9	26.8	−52.9	17.4	−62.3	40.8	−38.9
	Sup. (DeiT III)	400	**84.6**	**74.8**	**− 9.8**	**56.8**	**−27.8**	**43.3**	**−41.2**	**49.0**	**−35.5**	**56.1**	**−28.5**
	MAE	1600	73.0	60.7	−12.3	31.3	−41.6	20.7	−52.3	7.8	−65.1	20.7	−52.3
	CrossMAE	800	71.5	58.8	−12.7	31.8	−39.7	21.7	−49.8	7.4	−64.1	19.6	−51.9
	data2vec 2.0	200	70.5	58.6	**−11.9**	31.2	**−39.3**	23.7	**−46.8**	12.5	**−58.0**	**34.8**	**−35.7**
	I-JEPA	600	68.4	54.6	−13.8	20.6	−47.8	11.9	−56.5	4.4	−64.0	13.9	−54.5
	Bootleg	600	**77.5**	**65.1**	−12.4	**34.3**	−43.2	**24.4**	−53.1	**13.9**	−63.7	25.3	−52.2

For the attentive probes, Bootleg also has the highest performance on IN-1k and IN-V2, as shown in Table 8. However, IN-V2 is not shifted in domain from IN-1k so this does not demonstrate robustness to domain shift only the within-domain generalization capabilities of the model. Bootleg also performs best on IN-A and (for ViT-S and -B) IN-C, demonstrating it has the best robustness to these domain shifts across the SSL models. However, the Bootleg attentive probes perform poorly on IN-R and IN-Sketch, suggesting the attentive probe relies on features in the embeddings of the pretrained model that are not abstract enough to be robust to changes in artform/rendition style.

Table 8: **Attentive probe accuracy and robustness.** Performance of attentive probe fit on ImageNet-1k and evaluated on various robustness benchmarks, compared with cross-entropy supervised baselines. **Best** and second-best highlighted in each category.

Arch	Method	Ep.	IN-1k	IN-V2		IN-R		IN-Sketch		IN-A		IN-C	
ViT-S	Sup. (AugReg)	300	74.9	62.4	−12.5	33.2	−41.7	17.1	−57.8	8.8	−66.1	25.5	−49.4
	Sup. (DeiT III)	400	**81.3**	**70.8**	**−10.6**	**46.5**	**−34.8**	**35.0**	**−46.3**	**27.2**	**−54.2**	**42.7**	**−38.6**
	MAE	800	66.4	53.5	−13.0	28.3	−38.2	16.6	−49.8	3.5	−62.9	10.8	−55.7
	CrossMAE	800	68.8	56.8	**−12.0**	**31.8**	**−37.0**	19.9	**−48.9**	4.6	−64.1	14.3	−54.5
	data2vec 2.0	200	62.2	49.3	−12.8	24.0	−38.1	12.5	−49.6	2.7	−59.5	9.6	−52.6
	I-JEPA	600	61.9	49.0	−12.8	20.3	−41.5	8.8	−53.1	3.9	**−58.0**	9.5	**−52.3**
	Bootleg	600	**75.3**	**62.8**	−12.5	31.5	−43.8	**20.6**	−54.7	**10.6**	−64.7	**16.6**	−58.7
ViT-B	Sup. (torch)	300	81.0	69.6	−11.5	44.1	−37.0	29.4	−51.7	20.8	−60.2	39.5	−41.5
	Sup. (AugReg)	300	76.8	64.1	−12.6	37.4	−39.4	22.6	−54.2	11.3	−65.5	31.4	−45.3
	Sup. (DeiT III)	400	**83.7**	**73.6**	**−10.1**	**53.6**	**−30.1**	**40.1**	**−43.6**	**39.9**	**−43.8**	**50.7**	**−32.9**
	MAE	1600	76.0	63.8	−12.2	35.8	−40.2	22.8	−53.2	12.0	−64.0	16.6	−59.4
	CrossMAE	800	75.6	63.8	−11.8	**39.0**	**−36.5**	**25.4**	−50.2	11.9	−63.7	20.0	−55.6
	data2vec 2.0	200	73.7	61.3	−12.4	34.1	−39.6	24.8	**−48.9**	8.7	−65.0	19.3	**−54.4**
	I-JEPA	600	72.4	59.2	−13.2	25.7	−46.8	14.9	−57.6	8.7	−63.7	14.1	−58.3
	Bootleg	600	**79.2**	**67.5**	**−11.7**	36.2	−43.0	24.6	−54.6	**20.3**	**−58.9**	**22.3**	−56.9
ViT-L	Sup. (torch)	600	79.7	67.5	−12.2	40.8	−38.9	26.8	−52.9	17.4	−62.3	40.8	−38.9
	Sup. (DeiT III)	400	**84.6**	**74.8**	**− 9.8**	**56.8**	**−27.8**	**43.3**	**−41.2**	**49.0**	**−35.5**	**56.1**	**−28.5**
	MAE	1600	79.5	68.1	−11.4	**42.2**	−37.3	27.6	−51.9	21.8	−57.7	27.6	−51.9
	CrossMAE	800	78.7	67.2	−11.5	41.7	**−37.0**	28.6	−50.1	19.4	−59.3	25.5	−53.1
	data2vec 2.0	200	80.0	69.6	**−10.4**	41.1	−38.8	**30.8**	**−49.2**	28.9	**−51.1**	**42.5**	**−37.5**
	I-JEPA	600	72.3	58.8	−13.6	25.6	−46.7	14.5	−57.8	7.4	−64.9	16.0	−56.3
	Bootleg	600	**80.6**	**69.7**	−10.9	39.3	−41.3	27.8	−52.8	**29.0**	−51.6	28.6	−52.0

D TEST-TIME ADAPTATION WITH SAR

D.1 METHODOLOGY

We take the SSL-pretrained encoder and, while keeping it frozen, train linear and attentive probes atop of it to predict IN-1k classes (see Appx. A.1). We then evaluate how well the pretrained model can be adapted to new domain-shifted downstream datasets through test-time adaptation (TTA)

We use the SAR methodology of Niu et al. (2023). For each dataset, the pretrained encoder is adapted to the domain-shifted dataset exclusively by fine-tuning the affine transform parameters of the norm layers. We keep the final 3 blocks and the classifier head frozen. Performance is measured while adapting the model online using sharpness-aware entropy minimization. We use a batch size of 256; we sweep over 6 learning rate values—$\eta_{\text{max}} \in [1\times10^{-5}, 3\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}, 1\times10^{-3}, 3\times10^{-3}]$ per 64 samples—and select the best performing model.

We evaluate on IN-R and IN-C datasets. For IN-C, we perform the adaptation separately for each of the 15 corruption types (at severity 5) and average the performance before choosing the best LR.

D.2 APPLYING SAR TO FINE-TUNED MODELS

To accompany Table 3, we investigated the test-time adaptability of SSL pretrained models after they have been fine-tuned—is adaptability retained during the fine-tuning process?

Models were pretrained with SSL on IN-1k without labels, then fine-tuned with cross-entropy using the labels as described in Appx. A.2. We then used the methodology described in Appendix D to perform TTA of the fine-tuned models.

As shown in Table 9, we found all ViT-S models were much more adaptable to IN-C after fine-tuning than before it (Table 3). For ViT-B and L, we found data2vec 2.0 was much more adaptable and Bootleg less adaptable compared to the linear probe SAR results.

Table 9: **Test-time adaptation with SAR of fine-tuned models.** IN-C shows the mean accuracy over 15 corruptions. **Best** and second-best SSL method highlighted.

		IN-R			IN-C		
Arch	Method	Base	SAR	Δ	Base	SAR	Δ
ViT-S/16	Supervised (AugReg)	33.2	34.4	+1.1	25.5	44.8	**+19.3**
	Supervised (DeiT III)	**46.5**	**51.1**	**+4.5**	**42.7**	**47.8**	+5.1
	MAE	42.8	42.8	−0.0	32.7	41.8	**+9.1**
	CrossMAE	43.8	45.6	+1.8	34.6	43.0	+8.4
	data2vec 2.0	45.8	47.3	+1.5	36.9	45.7	+8.7
	I-JEPA	44.5	47.2	**+2.6**	38.8	42.6	+3.8
	Bootleg	**47.7**	**49.6**	+1.9	**39.9**	**47.5**	+7.6
ViT-B/16	Supervised (torch)	44.1	39.6	−4.4	39.5	49.1	+9.6
	Supervised (AugReg)	37.4	39.8	**+2.4**	31.4	53.4	**+22.0**
	Supervised (DeiT III)	**53.6**	**55.9**	+2.3	**50.7**	**61.0**	+10.3
	MAE	45.6	47.8	+2.2	42.4	47.0	+4.6
	CrossMAE	48.1	48.9	+0.8	44.3	**51.2**	**+6.9**
	data2vec 2.0	50.2	**53.2**	**+3.0**	45.6	49.5	+3.9
	I-JEPA	47.8	47.8	−0.0	**46.4**	44.8	−1.6
	Bootleg	**51.2**	52.1	+0.9	46.4	47.6	+1.3
ViT-L/16	Supervised (torch)	40.8	41.1	+0.3	40.8	45.4	+4.5
	Supervised (DeiT III)	**56.8**	**58.6**	**+1.8**	**56.1**	**63.7**	**+7.6**
	MAE	55.4	57.4	**+2.0**	51.0	53.6	+2.6
	CrossMAE	53.3	54.5	+1.2	50.7	51.3	+0.6
	data2vec 2.0	**59.1**	**61.0**	+2.0	**58.5**	**61.9**	**+3.5**
	I-JEPA	48.6	49.2	+0.6	46.2	46.7	+0.5
	Bootleg	57.3	57.4	+0.1	54.4	56.4	+2.0