
In-N-Out: Pre-Training and Self-Training using Auxiliary Information for Out-of-Distribution Robustness

Robbie Jones* Sang Michael Xie* Ananya Kumar*
Fereshte Khani Tengyu Ma Percy Liang

Stanford University

{rmjones, xie, ananya, fereshte, tengyuma, pliang}@cs.stanford.edu

Abstract

Many machine learning applications used to tackle climate change involve lots of unlabeled data (such as satellite imagery) along with auxiliary information such as climate data. In this work, we show how to use auxiliary information in a semi-supervised setting to improve both in-distribution and out-of-distribution (OOD) accuracies (e.g. for countries in Africa where we have very little labeled data). We show that 1) on real-world datasets, the common practice of using auxiliary information as additional input features improves in-distribution error but can hurt OOD. Oppositely, we find that 2) using auxiliary information as outputs of auxiliary tasks to pre-train a model improves OOD error. 3) To get the best of both worlds, we introduce In-N-Out, which first trains a model with auxiliary inputs and uses it to pseudolabel all the in-distribution inputs, then pre-trains a model on OOD auxiliary outputs and fine-tunes this model with the pseudolabels (self-training). We show both theoretically and empirically on remote sensing datasets for land cover prediction and cropland prediction that In-N-Out outperforms auxiliary inputs or outputs alone on both in-distribution and OOD error.

1 Introduction

When models are tested on distributions that are different from the training distribution, they typically suffer large drops in performance [44, 2, 22, 17, 3]. This is an especially salient obstacle in remote sensing applications, which includes central problems such as predicting poverty, crop type, and land cover from satellite imagery for downstream use in climate change mitigation [52, 21, 49, 40, 38]. In some developing African countries, labels are scarce due to the lack of economic resources to deploy human workers to conduct expensive surveys [21]. To make accurate predictions in these countries, we must extrapolate to out-of-distribution (OOD) examples across different geographic terrains and political borders to create a globally applicable model.

While labels are scarce, auxiliary information (e.g. climate data in remote sensing) is often available for every input. In many applications, we additionally have access to a large amount of unlabeled data (e.g. global satellite imagery) along with their corresponding auxiliary information. How should auxiliary information be used in this setting? One way is to use them directly as auxiliary input features (**aux-inputs**); another is to treat them as prediction outputs for an auxiliary task (**aux-outputs**) in pre-training or multi-task learning. Which approach is better for in-distribution or OOD performance?

In remote sensing, satellite imagery is paired with aux-inputs to predict the desired output [49, 54]. Aux-inputs can provide more features to improve in-distribution performance, but in this paper we find they can hurt OOD error. Conversely, aux-output methods such as pre-training, transfer learning, and multi-task learning may improve OOD performance by changing the inductive bias of the model through auxiliary supervision [50, 7]. We prove in a multi-task linear regression setting [13, 46] that aux-inputs can improve in-distribution error and worsen OOD error, while aux-outputs improve robustness to arbitrary covariate shift by first pre-training on unlabeled data to predict auxiliary information.

*Equal contribution.

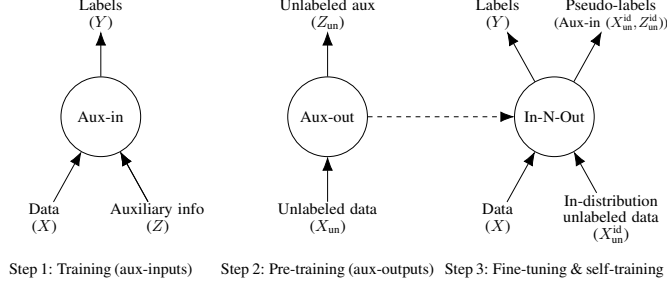


Figure 1: A sketch of the In-N-Out algorithm which consists of three steps: 1) use auxiliary information as input (Aux-in), 2) use auxiliary information as output targets in pretraining (Aux-out), 3) fine-tune the pretrained model from step 2 using the labeled data and in-distribution unlabeled data with pseudo-labels generated by the Aux-in model from step 1.

Can we do better than using auxiliary information as inputs or outputs alone? We propose In-N-Out to combine the benefits of auxiliary inputs and outputs (Figure 1). In-N-Out first uses an aux-inputs model, which has good in-distribution accuracy, to pseudolabel in-distribution unlabeled data. Then we use these pseudolabels for self-training, where a pretrained model (using aux-outputs) is fine-tuned on the larger labeled and pseudolabeled dataset. We prove that In-N-Out, which combines self-training and pretraining, further improves both in-distribution and OOD error over just pretraining.

We show empirical results on CelebA and two remote sensing tasks (land cover and cropland prediction) that matches the theory. On all datasets, In-N-Out improves OOD accuracy and has competitive or better in-distribution accuracy (1-2% in-distribution, 2-4% OOD over baseline on remote sensing tasks) over aux-inputs or aux-outputs alone.

2 Setup

Let $x \in \mathbb{R}^d$ be the input (e.g., a satellite image), $y \in \mathbb{R}$ be the target (e.g., crop type), and $z \in \mathbb{R}^T$ be auxiliary information—typically low-dimensional semantically meaningful features (e.g., soil type).

Training data. Let P_{id} and P_{ood} denote the underlying distribution of (x, y, z) triples in-distribution and out-of-distribution, respectively. The training data consists of (i) in-distribution labeled data $\{(x_i, y_i, z_i)\}_{i=1}^n \sim P_{\text{id}}$, (ii) in-distribution unlabeled data $\{(x_i^{\text{id}}, z_i^{\text{id}})\}_{i=1}^{m_1} \sim P_{\text{id}}$, and (iii) out-of-distribution unlabeled data $\{(x_i^{\text{ood}}, z_i^{\text{ood}})\}_{i=1}^{m_2} \sim P_{\text{ood}}$.

2.1 Models

We consider three common ways to use the auxiliary information (z) to learn a model.

Baseline (\hat{f}_{bs}). Baseline ignores auxiliary information and learns $\hat{f}_{\text{bs}}: x \mapsto y$ from labeled data.

Aux-inputs (\hat{f}_{in}). The aux-inputs model uses the auxiliary information as features and learns a function $\hat{f}_{\text{in}}: (x, z) \mapsto y$ from labeled data.

Aux-outputs (\hat{f}_{out}). The aux-outputs model leverages the auxiliary information z by using them as the prediction targets of T auxiliary tasks, in hopes that there is a low-dimensional feature representation w that is common to predicting both z and y . Training the aux-outputs model consists of two steps:

In the *pre-training* step, we use all the unlabeled data to learn a shared feature representation w . Specifically, we learn a feature map $\hat{h}_{\text{out}}: x \mapsto w$ and a mapping $\hat{g}_{\text{out}}^z: w \mapsto z$ such that $\hat{f}_{\text{out}} = \hat{g}_{\text{out}}^z \circ \hat{h}_{\text{out}}$ minimizes the training loss on all unlabeled data pairs (x, z) .

In the *transfer* step, the model uses the pre-trained feature map \hat{h}_{out} and the labeled data to learn the mapping $\hat{g}_{\text{out}}^y: w \mapsto y$ from feature representation to target. The estimate of the target mapping is $\hat{g}_{\text{out}}^y = \arg\min_{g_{\text{out}}^y} \hat{R}_{\text{trans}}(\hat{h}_{\text{out}}, g_{\text{out}}^y)$, where $\hat{R}_{\text{trans}}(\hat{h}_{\text{out}}, g_{\text{out}}^y) = \frac{1}{n} \sum_{i=1}^n \ell(g_{\text{out}}^y(\hat{h}_{\text{out}}(x_i)), y_i)$ is the *transfer empirical risk*. The final aux-outputs model is $\hat{f}_{\text{out}} = \hat{g}_{\text{out}}^y \circ \hat{h}_{\text{out}}$. Like the baseline model, the aux-outputs model ignores the auxiliary information for prediction.

3 In-N-Out: combining auxiliary inputs and outputs

We analyze these algorithms in a multi-task linear regression setting, formalized in Appendix B and C. We formalize all theorems in Appendix C and D and prove them in Appendix E, but give an outline here.

Algorithm 1: In-N-Out (see Section 3 for formal description)

Data: labeled data $\{(x_i, y_i, z_i)\}_{i=1}^n \sim P_{\text{id}}$, in-distribution
unlabeled data $\{(x_i^{\text{id}}, z_i^{\text{id}})\}_{i=1}^{m_1} \sim P_{\text{id}}$, OOD unlabeled data $\{(x_i^{\text{ood}}, z_i^{\text{ood}})\}_{i=1}^{m_2} \sim P_{\text{ood}}$
Learn aux-inputs model $\hat{f}_{\text{in}}: (x, z) \mapsto y$ from labeled in-distribution data $\{(x_i, y_i, z_i)\}_{i=1}^n \sim P_{\text{id}}$;
Pre-train $\hat{h}_{\text{out}}: x \mapsto z$ on aux-outputs from all unlabeled data $\{(x_i^{\text{id}}, z_i^{\text{id}})\}_{i=1}^{m_1} \cup \{(x_i^{\text{ood}}, z_i^{\text{ood}})\}_{i=1}^{m_2}$;
Return $\hat{f} = \hat{g} \circ \hat{h}_{\text{out}}: x \mapsto y$, trained with pseudolabels of $\hat{f}_{\text{in}}: \{(x_i, y_i)\}_{i=1}^n \cup \{(x_i^{\text{id}}, \hat{f}_{\text{in}}(x_i^{\text{id}}, z_i^{\text{id}}))\}_{i=1}^{m_1}$



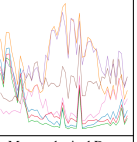
	CelebA	Cropland	Landcover
Visualization (x)			
Aux Info (z)	7 binary attributes	Vegetation, Lat/Lon	Meteorological Data
Target (y)	Male/female?	Cropland/not cropland?	Land cover class
ID-Split	People without hats	IA, MN, IL	Outside Africa
OOD-Split	People with hats	IN, KY	Africa

Figure 2: Summary of the datasets used in our experiments.

Aux-inputs help in-distribution, but hurt OOD. Proposition 1 in Appendix C shows that aux-inputs help in-distribution but Example 1 shows that *aux-inputs can hurt accuracy OOD*.

Aux-outputs help OOD. Our first main contribution is that *auxiliary outputs help under arbitrary covariate shift*, assuming we have lots of unlabeled data.

Theorem 1. *In the linear setting, the aux-outputs model improves the out-of-distribution risk R_{ood} over the baseline, for any OOD distribution:*

$$\mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{out}})] \leq \mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{bs}})]. \quad (1)$$

In-N-Out. We propose the In-N-Out algorithm (Algorithm 1), which combines both the aux-inputs and aux-outputs models for further complementary gains (Figure 1). Since the aux-inputs model has better in-distribution performance, we use it to pseudolabel in-distribution unlabeled data (e.g., data from non-African locations). The pseudolabeled data provides more effective training samples (self-training) to fine-tune a model pre-trained by predicting aux-outputs on all unlabeled data (e.g., including Africa). The In-N-Out model $\hat{f} = \hat{g} \circ \hat{h}_{\text{out}}$ optimizes the empirical risk on labeled and pseudolabeled data:

$$\hat{g} = \underset{g}{\operatorname{argmin}} (1 - \lambda) \hat{R}_{\text{trans}}(\hat{h}_{\text{out}}, g) + \lambda \hat{R}_{\text{st}}(g) \quad (2)$$

where $\hat{R}_{\text{st}}(g) = \frac{1}{m_1} \sum_{i=1}^{m_1} \ell(g(\hat{h}_{\text{out}}(x_i^{\text{id}})), \hat{f}_{\text{in}}(x_i^{\text{id}}, z_i^{\text{id}}))$ is the self-training loss on pseudolabels from the aux-inputs model, and $\lambda \in [0, 1]$ trades off between labeled and pseudolabeled losses. In our experiments, we fine-tune \hat{g} and \hat{h}_{out} together. In the linear regression setting, we show that In-N-Out improves over aux-outputs for arbitrary covariate shifts. The formal result is in Theorem 4 in Appendix D, with the proof in Appendix E.

Theorem 2 (Informal). *In the linear setting, if X and Z together contain most of the information required to predict Y ($\mathbb{E}[\text{Var}(Y | X, Z)]$ is small), the In-N-Out model improves the out-of-distribution risk R_{ood} over aux-outputs, for any OOD distribution:*

$$\mathbb{E}[R_{\text{ood}}(\hat{f})] < \mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{out}})]. \quad (3)$$

4 Experiments

We show on remote sensing datasets for land cover and cropland prediction that aux-inputs can hurt OOD performance while aux-outputs improve OOD performance. In-N-Out improves OOD accuracy and has competitive or better in-distribution accuracy over other models on all datasets.

4.1 Datasets and Experimental Setup

Cropland. Crop type or cropland prediction is an important intermediate problem for crop yield prediction, which can aid efficient agricultural practices (“precision agriculture”) that are adaptable to different crop regions [5, 23, 28, 38]. The input x is a 50×50 RGB image taken by a satellite, the target

	CelebA		Cropland		Landcover	
	ID Acc	OOD Acc	ID Acc	OOD Acc	ID Acc	OOD Acc
Baseline	90.46 \pm 0.85	72.64 \pm 1.39	94.50 \pm 0.11	90.30 \pm 0.75	75.92 \pm 0.25	58.31 \pm 1.87
Aux-inputs	92.36 \pm 0.29	77.4 \pm 1.33	95.34 \pm 0.22	84.15 \pm 4.23	76.58 \pm 0.44	54.78 \pm 2.01
Aux-outputs	94.0 \pm 0.24	77.68 \pm 0.59	95.12 \pm 0.15	91.63 \pm 0.21	72.48 \pm 0.37	61.03 \pm 0.97
In-N-Out (no pretrain)	93.8 \pm 0.56	78.54 \pm 1.31	94.93 \pm 0.15	91.23 \pm 0.61	76.54 \pm 0.23	59.19 \pm 0.98
In-N-Out	93.42 \pm 0.36	79.42 \pm 0.70	95.45 \pm 0.16	91.94 \pm 0.57	77.43 \pm 0.39	61.53 \pm 0.74
In-N-Out + repeated ST	93.76 \pm 0.46	80.38 \pm 0.68	95.53 \pm 0.19	92.18 \pm 0.40	77.10 \pm 0.30	62.61 \pm 0.58

Table 1: Accuracy (%) of models using auxiliary information as input, output, or both. In-N-Out improves ID and OOD over aux-inputs or aux-outputs. Results are averaged over 5 trials with 90% intervals. Repeated ST refers to one round of self-training on top of In-N-Out.

y is a binary label that is 1 when the image contains majority cropland, and the auxiliary information z is the center location coordinate plus 50×50 vegetation-related bands. We use the Cropland dataset from Wang et al. [49], with data from the US Midwest. We designate Iowa, Missouri, and Illinois as in-distribution and Indiana and Kentucky as OOD. Following Wang et al. [49], we use a U-Net-based model [39]. See Appendix F.1 for details.

Landcover. Land cover prediction involves classifying the land cover type (e.g., “grasslands”) from satellite data at a location [15, 40]). These land cover models can aid climate scientists in tracking ecological deterioration such as deforestation or melting ice sheets on a global scale. The input x is a time series measured by NASA’s MODIS satellite [48], the target y is one of 6 land cover classes, and the auxiliary information z is climate data (e.g. temperature) from the ERA5 satellite [4]. We designate non-African locations as in-distribution and Africa as OOD. We use a 1D-CNN to handle the temporal structure in the MODIS data. See Appendix F.2 for details.

CelebA. In CelebA, the input x is a RGB image (resized to 64×64), the target y is a binary label for gender, and the auxiliary information z are 7 (of 40) binary-valued attributes (e.g. presence of makeup, beard). We designate the set of images where the celebrity is wearing a hat as OOD. We use a ResNet18 as the backbone model architecture for all models (see Appendix F.3 for details).

4.2 Results

Table 1 compares the in-distribution (ID) and OOD accuracy of different methods. Each method is trained with early-stopping and hyperparameters are chosen using a validation set. In our experiments, we also consider augmenting In-N-Out models with repeated self-training (repeated ST), which has fueled recent improvements in both domain adaptation and ImageNet classification [42, 53]. For one additional round of repeated ST, we use the In-N-Out model to pseudolabel all unlabeled data (both ID and OOD) and also initialize the weights with the In-N-Out model. In all datasets, pretraining with aux-outputs improves OOD performance over the baseline, and In-N-Out (with or without repeated ST) generally improves both in- and out-of-distribution performance over all other models.

In our remote sensing datasets, aux-inputs can induce a tradeoff where increasing ID accuracy hurts OOD performance. In cropland prediction, even with a small geographic shift (US Midwest), the baseline model has a significant drop from ID to OOD accuracy (4%). The aux-inputs model improves ID accuracy almost 1% above the baseline but OOD accuracy drops 6%. In land cover prediction, using climate features as aux-inputs decreases OOD accuracy by over 4% compared to the baseline. The aux-outputs model improves OOD, but decreases ID accuracy by 3%. In both datasets, In-N-Out-based models generally improve both in- and OOD accuracy over all models.

In CelebA, using auxiliary information either as aux-inputs or outputs improves both ID (2-4%) and OOD accuracy (5%). In-N-Out achieves the best OOD performance and comparable ID performance even though there is no tradeoff between ID and OOD accuracy. The discrepancy in ID and OOD behavior compared to Cropland and Landcover underscores the importance of using real-world remote sensing datasets in addition to synthetic datasets.

5 Conclusion

While auxiliary inputs improve in-distribution and OOD on standard curated datasets, they can hurt OOD on real-world tasks important for addressing climate change. In contrast, using auxiliary information as outputs by pretraining improves OOD performance. In-N-Out combines the strengths of auxiliary inputs and outputs for further improvements. We note that the division between inputs and auxiliary information is not well-defined. Our framework applies generally to any division of the features, but an important further question is how to optimally choose what to use as auxiliary information under distribution shifts.

Acknowledgments and Disclosure of Funding

We thank Sherrie Wang, Andreas Schlueter, Albert Gu, Daniel Levy, Pang Wei Koh, and Shiori Sagawa, and anonymous reviewers for their valuable help and comments. This work was supported by an Open Philanthropy Project Award, an NSF Frontier Award as part of the Center for Trustworthy Machine Learning (CTML), SDSI, and SAIL at Stanford University. SMX was supported by an NDSEG Fellowship. AK was supported by a Stanford Graduate Fellowship.

References

- [1] S. Ahmad, A. Kalra, and H. Stephen. Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources*, 33(1):69–80, 2010.
- [2] E. AlBadawy, A. Saha, and M. Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med Phys.*, 45, 2018.
- [3] J. Blitzer and F. Pereira. Domain adaptation of natural language processing systems. *University of Pennsylvania*, 2007.
- [4] C3S. ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, 2017.
- [5] Y. Cai, K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow, and Z. Li. A high-performance and in-season classification system of field-level crop types using time-series landsat data and a machine learning approach. *Remote Sensing of Environment*, 210:74–84, 2018.
- [6] Y. Carmon, A. Ragunathan, L. Schmidt, P. Liang, and J. C. Duchi. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [7] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.
- [8] R. Caruana and V. R. de Sa. Benefitting from the variables that variable selection discards. *Journal of Machine Learning Research (JMLR)*, 3, 2003.
- [9] Y. Chen, C. Wei, A. Kumar, and T. Ma. Self-training avoids using spurious features under domain shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [10] R. DeFries, M. Hansen, and J. Townshend. Global discrimination of land cover types from metrics derived from AVHRR pathfinder data. *Remote Sensing of Environment*, 54(3):209–222, 1995.
- [11] R. S. DeFries and J. Townshend. NDVI-derived land cover classifications at a global scale. *International Journal of Remote Sensing*, 15(17):3567–3586, 1994.
- [12] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics (ACL)*, pages 4171–4186, 2019.
- [13] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv*, 2020.
- [14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17, 2016.
- [15] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.
- [16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [17] D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning (ICML)*, 2019.
- [18] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

- [19] J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.
- [20] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Conference on Learning Theory (COLT)*, 2012.
- [21] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353, 2016.
- [22] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.
- [23] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard. Crop yield forecasting on the canadian prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology*, 218:74–84, 2016.
- [24] S. K and Z. A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.
- [25] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman. 1d convolutional neural networks and applications: A survey. *arXiv preprint arXiv:1905.03554*, 2019.
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.
- [27] A. Kumar, T. Ma, and P. Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning (ICML)*, 2020.
- [28] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017.
- [29] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, 2016.
- [30] A. Li, S. Liang, A. Wang, and J. Qin. Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. *Photogrammetric Engineering & Remote Sensing*, 73(10):1149–1157, 2007.
- [31] R. Lunetta, J. F. Knight, J. E. J. G. Lyon, and L. D. Worthy. Land-cover change detection using multi-temporal MODIS NDVI data. *Remote sensing of environment*, 105(2):142–154, 2006.
- [32] A. E. Maxwell, T. A. Warner, and F. Fang. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, 39(9):2784–2817, 2018.
- [33] A. Najafi, S. Maeda, M. Koyama, and T. Miyato. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [34] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2020.
- [35] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. In *Very Large Data Bases (VLDB)*, number 3, pages 269–282, 2017.
- [36] A. J. Ratner, C. M. D. Sa, S. Wu, D. Selsam, and C. Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3567–3575, 2016.
- [37] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

- [38] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. R., N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019.
- [39] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *arXiv*, 2015.
- [40] M. Rußwurm, S. Wang, M. Korner, and D. Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 200–201, 2020.
- [41] S. Santurkar, D. Tsipras, and A. Madry. Breeds: Benchmarks for subpopulation shift. *arXiv*, 2020.
- [42] R. Shu, H. H. Bui, H. Narui, and S. Ermon. A DIRT-T approach to unsupervised domain adaptation. In *International Conference on Learning Representations (ICLR)*, 2018.
- [43] M. Sugiyama, M. Krauledat, and K. Muller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research (JMLR)*, 8:985–1005, 2007.
- [44] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.
- [45] T. Tao. *Topics in random matrix theory*. American Mathematical Society, 2012.
- [46] N. Tripuraneni, M. I. Jordan, and C. Jin. On the theory of transfer learning: The importance of task diversity. *arXiv*, 2020.
- [47] J. Uesato, J. Alayrac, P. Huang, R. Stanforth, A. Fawzi, and P. Kohli. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [48] E. Vermote. MOD09A1 MODIS/terra surface reflectance 8-day L3 global 500m SIN grid V006. <https://doi.org/10.5067/MODIS/MOD09A1.006>, 2015.
- [49] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12, 2020.
- [50] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3, 2016.
- [51] S. Wu, H. R. Zhang, and C. Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2020.
- [52] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.
- [53] Q. Xie, M. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. *arXiv*, 2020.
- [54] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11, 2020.
- [55] B. Zoph, G. Ghiasi, T. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. V. Le. Rethinking pre-training and self-training. *arXiv*, 2020.

A Additional related works

Multi-task learning and weak supervision. Caruana and de Sa [8] proposed using poor input features as a multi-task output, but do not theoretically analyze this. Wu et al. [51] also study multi-task linear regression. However, their auxiliary tasks must have true parameters that are closely aligned (small cosine distance) to the target task. Similarly, weak supervision works assume access to weak labels correlated with the true label [36, 35]. In our paper, we make no assumptions about the alignment of the auxiliary and target tasks beyond a shared latent variable while also considering distribution shifts.

Transfer learning, pre-training, and self-supervision. We support empirical works that show the success of transfer learning and pre-training in vision and NLP [26, 24, 12]. Theoretically, Du et al. [13], Tripuraneni et al. [46] study pre-training in a similar linear regression setup. They show in-distribution generalization bound improvements, but do not consider OOD robustness or combining with auxiliary inputs. Hendrycks et al. [18] shows empirically that self-supervision can improve robustness to synthetic corruptions. We show similar robustness benefits for pre-training on auxiliary information (not part of the original input).

Self-training for robustness. [34] analyze robust self-training (RST) Carmon et al. [6], Najafi et al. [33], Uesato et al. [47], which improves the tradeoff between standard and adversarially robust accuracy, in min-norm linear regression. While related, we work in multi-task linear regression, study pre-training, and prove robustness to arbitrary covariate shifts (rather than adversarial perturbations). Kumar et al. [27] show that repeated self-training on gradually shifting unlabeled data can enable adaptation over time. In-N-Out is complementary and may provide better pseudolabels in each step of this method. Chen et al. [9] show that self-training can remove spurious features for Gaussian input features in linear models, whereas our results hold for general input distributions (with density). Zoph et al. [55] show that self-training and pre-training combine for in-distribution gains. We provide theory to support this and also show benefits for OOD robustness.

Domain adaptation. Domain adaptation works account for covariate shift by using unlabeled data from a target domain to adapt the model [42, 19, 14]. Often, these methods [42, 19] have a self-training or entropy minimization component that benefits from having a better model in the target domain to begin with. Similarly, domain adversarial methods [14] rely on the inductive bias of the source-only model to correctly align the source and target distributions. In-N-Out may provide a better starting point for these domain adaptation methods.

B Formal Setup

Let $x \in \mathbb{R}^d$ be the input (e.g., a satellite image), $y \in \mathbb{R}$ be the target (e.g., crop type), and $z \in \mathbb{R}^T$ be auxiliary information—typically low-dimensional semantically meaningful features (e.g., soil type).

Training data. Let P_{id} and P_{ood} denote the underlying distribution of (x, y, z) triples in-distribution and out-of-distribution, respectively. The training data consists of (i) in-distribution labeled data $\{(x_i, y_i, z_i)\}_{i=1}^n \sim P_{\text{id}}$, (ii) in-distribution unlabeled data $\{(x_i^{\text{id}}, z_i^{\text{id}})\}_{i=1}^{m_1} \sim P_{\text{id}}$, and (iii) out-of-distribution unlabeled data $\{(x_i^{\text{ood}}, z_i^{\text{ood}})\}_{i=1}^{m_2} \sim P_{\text{ood}}$.

Loss metrics. Our goal is to learn a model from input and auxiliary information to the target, $f : \mathbb{R}^d \times \mathbb{R}^T \rightarrow \mathbb{R}$. For a loss function ℓ , the in-distribution population risk of the model f is $R_{\text{id}}(f) = \mathbb{E}_{x, y, z \sim P_{\text{id}}}[\ell(f(x, z), y)]$, and its OOD population risk is $R_{\text{ood}}(f) = \mathbb{E}_{x, y, z \sim P_{\text{ood}}}[\ell(f(x, z), y)]$.

B.1 Models

We consider three common ways to use the auxiliary information (z) to learn a model.

Baseline. The baseline minimizes the empirical risk on labeled data while ignoring the auxiliary information (accomplished by setting z to 0):

$$\hat{f}_{\text{bs}} = \underset{f}{\operatorname{argmin}} \sum_{i=1}^n \ell(f(x_i, 0), y_i). \quad (4)$$

Aux-inputs. The aux-inputs model minimizes the empirical risk on labeled data while using the auxiliary information as features:

$$\hat{f}_{\text{in}} = \underset{f}{\operatorname{argmin}} \sum_{i=1}^n \ell(f(x_i, z_i), y_i). \quad (5)$$

Aux-outputs. The aux-outputs model leverages the auxiliary information z by using them as the prediction targets of T auxiliary tasks, in hopes that there is a low-dimensional feature representation that is common to predicting both z and y . Training the aux-outputs model consists of two steps:

In the *pre-training* step, we use all the unlabeled data to learn a shared feature representation. Let $h: \mathbb{R}^d \rightarrow \mathbb{R}^k$ denote a feature map and $g_{\text{out}}^z: \mathbb{R}^k \rightarrow \mathbb{R}^T$ denote a mapping from feature representation to the auxiliary outputs. Let ℓ_{aux} denote the loss function for the auxiliary information. We define the empirical risk of h and g_{out}^z as:

$$\hat{R}_{\text{pre}}(h, g_{\text{out}}^z) = \frac{1}{m_1 + m_2} \left(\sum_{i=1}^{m_1} \ell_{\text{aux}}(g_{\text{out}}^z(h(x_i^{\text{id}})), z_i^{\text{id}}) + \sum_{i=1}^{m_2} \ell_{\text{aux}}(g_{\text{out}}^z(h(x_i^{\text{ood}})), z_i^{\text{ood}}) \right). \quad (6)$$

The estimate of the feature map is $\hat{h}_{\text{out}} = \arg\min_h \min_{g_{\text{out}}^z} \hat{R}_{\text{pre}}(h, g_{\text{out}}^z)$.

In the *transfer* step, the model uses the pre-trained feature map \hat{h}_{out} and the labeled data to learn the mapping $g_{\text{out}}^y: \mathbb{R}^k \rightarrow \mathbb{R}$ from feature representation to target y . We define the transfer empirical risk as:

$$\hat{R}_{\text{trans}}(\hat{h}_{\text{out}}, g_{\text{out}}^y) = \frac{1}{n} \sum_{i=1}^n \ell(g_{\text{out}}^y(\hat{h}_{\text{out}}(x_i)), y_i) \quad (7)$$

The estimate of the target mapping is $\hat{g}_{\text{out}}^y = \arg\min_{g_{\text{out}}^y} \hat{R}_{\text{trans}}(\hat{h}_{\text{out}}, g_{\text{out}}^y)$. The final aux-outputs model is

$$\hat{f}_{\text{out}}(x, z) = \hat{g}_{\text{out}}^y(\hat{h}_{\text{out}}(x)). \quad (8)$$

Like the baseline model, the aux-outputs model ignores the auxiliary information for prediction.

C Theoretical Analysis of Aux-inputs and Aux-outputs Models

We now analyze the baseline, aux-inputs, and aux-outputs models introduced in Section 2. Our setup extends a linear regression setting commonly used for analyzing multi-task problems [13, 46].

Setup. See Figure 3 for the graphical model. Let $w = B^*x \in \mathbb{R}^k$ be a low-dimensional latent feature ($k \leq d$) shared between auxiliary information z and the target y . Let $u \in \mathbb{R}^m$ denote unobserved latent variables not captured in x . We assume z and y are linear functions of u and w :

$$y = \theta_w^\top w + \theta_u^\top u + \epsilon, \quad (9)$$

$$z = A^*w + C^*u, \quad (10)$$

where $\epsilon \sim P_\epsilon$ denotes noise with mean 0 and variance σ^2 . As in [13], we assume $T \geq k$ and $T \geq m$ and A^*, B^* and C^* have rank k .

Data. Let P_x and P_u denote the underlying distribution of x and u in-distribution, and let P'_x, P'_u denote their underlying distribution OOD. We assume x and u are independent, have bounded density everywhere, and have finite invertible covariance matrices. We assume the mean of u is zero in- and out-of-distribution². We assume we have $n \geq m + d$ in-distribution labeled training examples and unlimited access to unlabeled data both in- and out-of-distribution, a common assumption in unsupervised domain adaptation [43, 27, 34].

Loss metrics. We use squared-error for target and auxiliary losses: $\ell(\hat{y}, y) = (y - \hat{y})^2$ and $\ell_{\text{aux}}(z, z') = \|z - z'\|_2^2$.

Models. We assume all model families $(f, h, g_{\text{out}}^z, g_{\text{out}}^y)$ in Section 2 are linear.

Let $\mathcal{P} = (A^*, B^*, C^*, \theta_w, \theta_u, P_x, P_u)$ denote the problem setting which satisfies all the assumptions.

C.1 Auxiliary inputs help in-distribution, but can hurt OOD

We show that the aux-inputs model (5) performs better than the baseline model (4) in-distribution but might perform worse than the baseline model OOD. Intuitively, the target y depends on both the inputs x (through w) and latent variable u (Figure Figure 3). The baseline model only uses x to predict y ; thus it cannot capture the variation in y due to u . On the other hand, the aux-inputs model uses x and z to predict y . Since z is a function of x (through w) and u , u can be recovered from x and z by inverting this relation. The aux-inputs model can then combine u and x to predict y better.

Let $\sigma_u^2 = \mathbb{E}_{u \sim P_u}[(\theta_u^\top u)^2]$ denote the (in-distribution) variance of y due to the latent variables u . The following proposition shows that if $\sigma_u^2 > 0$ then with enough training examples the aux-inputs model has lower in-distribution population risk than the baseline model.³

²This is not limiting because bias in z can be folded into x .

³Since z is typically low-dimensional and x is high-dimensional (e.g. images), we only need a slightly larger number of examples for the aux-inputs model before it outperforms the baseline.

Proposition 1. For all problem settings \mathcal{P} , P_ϵ , assuming regularity conditions (bounded x , u , sub-Gaussian noise ϵ , and $T = m$), and $\sigma_u^2 > 0$, for all $\delta > 0$, there exists N such that for $n \geq N$ number of training points, with probability at least $1 - \delta$ over the training examples, the aux-inputs model improves over the baseline:

$$R_{id}(\hat{f}_{in}) < R_{id}(\hat{f}_{bs}) \quad (11)$$

Although using z as input leads to better in-distribution performance, we show that the aux-inputs model can perform worse than the baseline model in OOD for any number of training examples. The aux-inputs model learns to predict $\hat{y} = \hat{\theta}_{x,in}^\top x + \hat{\theta}_{z,in}^\top z$, where $\hat{\theta}_{z,in}$ is an approximation to the true parameter θ_z , that has some error. Out-of-distribution u and hence z might have much higher variance than x , which would magnify the error $\hat{\theta}_{z,in} - \theta_z$ and lead to worse predictions.

Example 1. There exists a problem setting \mathcal{P} , P_ϵ , such that for every n , there is some test distribution P'_x, P'_u with:

$$\mathbb{E}[R_{ood}(\hat{f}_{in})] > \mathbb{E}[R_{ood}(\hat{f}_{bs})] \quad (12)$$

C.2 Pre-training improves risk under arbitrary covariate shift

While using z as inputs can worsen performance relative to the baseline, our first main result is that the aux-outputs model outperforms the baseline model under arbitrary covariate shifts.

Theorem 3. For all problem settings \mathcal{P} , P_ϵ , and for all test distributions P'_x and P'_u :

$$\mathbb{E}[R_{ood}(\hat{f}_{out})] \leq \mathbb{E}[R_{ood}(\hat{f}_{bs})] \quad (13)$$

See Appendix E for full proof. Intuitively, pre-training by learning a low rank linear model from x to z allows us to learn the lower dimensional feature space $w = B^*x$ (up to symmetries). The aux-outputs model learns a linear map from the lower-dimensional w to y , while the baseline predicts y directly from x . Without distribution shift, standard techniques show that the aux-outputs model has better risk since w is lower dimensional than x . In particular, the in-domain risk only depends on the dimension but not on the conditioning of the data. In contrast, the worst case risk under distribution shift depends on the conditioning of the data, which could be worse for w than x . Our proof shows that the worst case risk (over all x and u) is still better for the aux-outputs model by “zeroing-out” error directions when projecting to the low-dimensional feature representation.

D In-N-Out improves risk under arbitrary covariate shift

Setup. For the theory we analyze a slightly modified version of In-N-Out where the aux-inputs model is trained on the features $\hat{h}_{out}(x)$ and we self-train on population unlabeled data. We train an aux-inputs model $\hat{g}_{in} : \mathbb{R}^k \times \mathbb{R}^T \rightarrow \mathbb{R}$ given by $\hat{g}_{in} = \arg\min_g \frac{1}{n} \sum_{i=1}^n \ell(g(\hat{h}_{out}(x_i), z_i), y_i)$. The self-training loss is: $R_{st}(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{z \sim P_{z|x_i}} [\ell(g(\hat{h}_{out}(x_i)), \hat{g}_{in}(\hat{h}_{out}(x_i), z_i))]$, and we learn $\hat{g} = \arg\min_g R_{st}(g)$. For input x, z the model outputs $\hat{g}(\hat{h}_{out}(x))$. For the theory, we assume all model families are linear.

We show that In-N-Out helps on top of pre-training, as long as the auxiliary features give us lots of information about y relative to the noise ϵ in-distribution—in particular, if σ_u^2 is much larger than σ^2 .

Theorem 4. In the linear setting, for all problem settings \mathcal{P} with $\sigma_u^2 > 0$, test distributions P'_x, P'_u , and $\delta > 0$, there exists $a, b > 0$ such that for all P_ϵ , with probability at least $1 - \delta$ over the training examples and test example $x' \sim P'_x$, the ratio of the excess risks (for all σ^2 small enough that $a - b\sigma^2 > 0$) is:

$$\frac{R_{in-out}^{ood} - R^*}{R_{out}^{ood} - R^*} \leq \frac{\sigma^2}{a - b\sigma^2} \quad (14)$$

Here $R^* = \min_{f^*} R_{ood}(f^*)$ is the minimum possible (Bayes-optimal) OOD risk, $R_{in-out}^{ood} = \mathbb{E}_{y' \sim P'_{y'|x'}} [\ell(\hat{g}(\hat{h}_{out}(x')), y')]$ is the risk of the In-N-Out model on test example x' , and $R_{out}^{ood} = \mathbb{E}_{y' \sim P'_{y'|x'}} [\ell(\hat{g}_{out}^y(\hat{h}_{out}(x')), y')]$ is the risk of the aux-outputs model on test example x' .

Remark 1. As $\sigma \rightarrow 0$, the excess risk ratio of In-N-Out to Aux-outputs goes to 0, and the In-N-Out estimator is much better than the aux-outputs estimator.

The proof of the result is in Appendix E, but we give high level intuition here. Since u can be recovered from w and z , we can write $y = \gamma_w^\top w + \gamma_z^\top z + \epsilon$. We train an aux-inputs model \hat{g}_{in} from w, z to y on finite labeled data—since the noise $\sigma^2 = \mathbb{E}[\epsilon^2]$ is small this model is very accurate. In the special case where $\epsilon = 0$, \hat{g}_{in} predicts y perfectly from w, z . For each training example w_i , we sample many $z \sim P_{z|w_i}$ and pseudolabel the w_i, z examples (very accurately). We then minimize the mean-squared

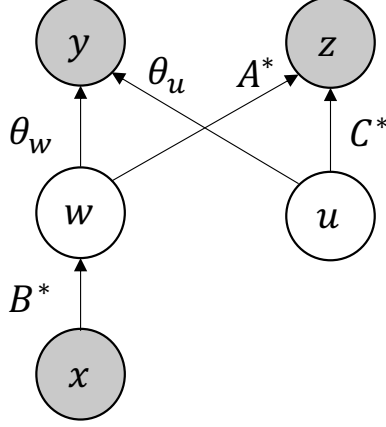


Figure 3: Graphical model for our theoretical setting, where auxiliary information z is related to targets y through the latent variable w and latent noise u .

error, which is equivalent to predicting the mean \hat{y}_i of these pseudolabels from w_i . This essentially averages over latent noise u , so \hat{y}_i is a denoised version of y_i . In the special case where $\epsilon = 0$, the pseudolabel $\hat{y}_i = \theta_w^\top w$ (without the u term) so we learn the Bayes-opt model θ_w .

The technical challenge is proving that self-training helps under arbitrary covariate shift even when $\epsilon > 0$ (the aux-inputs model is very accurate but not perfect). The proof reduces to showing that the max singular value for the In-N-Out error matrix is less than the min-singular value of the aux-outputs error matrix with high probability. A core part of the argument is to lower bound the min-singular value of a random matrix (Lemma 4). This uses techniques from random matrix theory (see e.g Chapter 2.7 in Tao [45]), the high level idea is to show that with probability $1 - \delta$ each column of the random matrix has a (not too small) component orthogonal to all other columns.

E Proof for Sections C and 3

Our theoretical setting assumes all the model families are linear. We begin by specializing the setup in Section 2 and defining all the necessary matrices. A word on notation: if unspecified, expectations are taken over all random variables.

Data matrices: We have finite labeled data in-distribution: $n \geq d + m$ input examples $X \in \mathbb{R}^{n \times d}$, where each row $X_i \sim P_x$ is an example sampled independently. We have an *unobserved* latent matrix: $U \in \mathbb{R}^{n \times m}$ where each row $U_i \sim P_u$ is sampled independently. U is not used by any of the models, but we will reference U in our analysis. We have labels $Y \in \mathbb{R}^n$ and auxiliary data $Z \in \mathbb{R}^{n \times T}$, where each row Y_i, Z_i is sampled jointly given input example X_i, U_i , that is: $Y_i, Z_i \sim P_{y,z|X_i, U_i}$.

E.1 Models and evaluation

Baseline: ordinary least squares estimator that uses x only, so $\hat{\theta}_{x,ols} = \text{argmin}_{\theta'} \|Y - X\theta'\|_2$. Given a test example x, z , the baseline method predicts $\hat{f}_{bs}(x, z) = \hat{\theta}_{x,ols}^\top x$, ignoring z . In closed form, $\hat{\theta}_{x,ols} = (X^\top X)^{-1} X^\top Y$.

Aux-inputs: least squares estimator using x and auxiliary z as input: $\hat{\theta}_{x,in}, \hat{\theta}_{z,in} = \text{argmin}_{\theta'_x, \theta'_z} \|Y - (X\theta'_x + Z\theta'_z)\|_2$. The input method predicts $\hat{\theta}_{x,in}^\top x + \hat{\theta}_{z,in}^\top z$ for a test example x, z . In closed form, letting $X_Z = [X; Z]$, where we append the columns so that $X_Z \in \mathbb{R}^{n \times (d+T)}$, $[\hat{\theta}_{x,in}, \hat{\theta}_{z,in}] = (X_Z^\top X_Z)^{-1} X_Z^\top Y$.

Aux-outputs: pretrains on predicting z from x on unlabeled data to learn a mapping from x to w , then learns a regression model on top of this latent embedding w . In the *pretraining* step: use unlabeled data to learn the feature-space embedding \hat{B} :

$$\hat{A}, \hat{B} = \text{argmin}_{A, B} \mathbb{E}_{x \sim P_x} [\|ABx - z\|_2^2] \quad A \in \mathbb{R}^{T \times k}, B \in \mathbb{R}^{k \times d} \quad (15)$$

The *transfer* step solves a lower dimensional regression problem from w to y : $\hat{\theta}_{w,out} = \text{argmin}_{\theta'_w} \|Y - X\hat{B}^\top \theta'_w\|_2$. Given a test example x , the output model predicts $\hat{\theta}_{w,out}^\top \hat{B}x$.

In-N-Out: First learn an output model \hat{A}, \hat{B} , and let $W = X \hat{B}^\top$ be the feature matrix. Next, train an input model on the feature space w so we get:

$$\hat{\gamma}_w, \hat{\gamma}_z = \underset{\gamma_w, \gamma_z}{\operatorname{argmin}} \| (Y - (W\gamma_w + Z\gamma_z))^2 \| \quad (16)$$

We now use the input model to pseudolabel our in-domain unlabeled examples, and self-train a model *without* z on these pseudolabels. Given each point w , we produce a pseudolabel $\mathbb{E}_{z \sim P_{z|w}} [\hat{\gamma}_w^\top w + \hat{\gamma}_z^\top z] = (\hat{\gamma}_w + A^\top \hat{\gamma}_z)^\top w$. We now learn a least squares estimator from w to the pseudolabels which gives us the In-N-Out estimator $\hat{\theta}_w$:

$$\hat{\theta}_w = \hat{\gamma}_w + A^\top \hat{\gamma}_z \quad (17)$$

Note that we do not actually have access to A , this is just the closed form for the final estimator that self-training on the pseudolabels gives us. Given a test example x , In-N-Out predicts $\hat{\theta}_w^\top \hat{B}x$.

E.2 Auxiliary inputs help in-distribution

The proof of Proposition 1 is fairly standard. We first give a brief sketch, specify the additional regularity conditions, and then give the proof. We lower bound the risk of the baseline by $\sigma_u^2 + \sigma^2$ since this is the Bayes-opt risk of using only x but not z to predict y . We upper bound the risk of the aux-inputs model which uses x, z to predict y , which is the same as upper bounding the risk in random design linear regression. For this upper bound we use Theorem 1 in Hsu et al. [20] (note that there are multiple versions of this paper, and we specifically use the Arxiv version, e.g. available at <https://arxiv.org/abs/1106.2363>). As such, we inherit their regularity conditions. In particular, we assume:

1. x, u are upper bounded almost surely. This is a technical condition, and can be replaced with sub-Gaussian tail assumptions [20].
2. The noise ϵ is sub-Gaussian with variance parameter σ^2 .
3. The latent dimension m and auxiliary dimension T are equal so that the inputs to the aux-inputs model have invertible covariance matrix.⁴

Restatement of Proposition 1. *For all problem settings \mathcal{P} , P_ϵ , assuming regularity conditions (bounded x, u , sub-Gaussian noise ϵ , and $T = m$), and $\sigma_u^2 > 0$, for all $\delta > 0$, there exists N such that for $n \geq N$ number of training points, with probability at least $1 - \delta$ over the training examples, the aux-inputs model improves over the baseline:*

$$R_{id}(\hat{f}_{in}) < R_{id}(\hat{f}_{bs}) \quad (18)$$

Proof. Lower bound risk of baseline: First, we lower bound the expected risk of the baseline by $\sigma_u^2 + \sigma^2$. Intuitively, this is the irreducible error—no linear classifier using only x can get better risk than $\sigma_u^2 + \sigma^2$ because of intrinsic noise in the output y . Let $\theta_x = B^{\star\top} \theta_w$ be the optimal baseline parameters. We have:

$$R_{id}(\hat{f}_{bs}) = \mathbb{E}_{x, y, z \sim P_{id}} [(y - \hat{\theta}_{x,ols}^\top x)^2] \quad (19)$$

$$= \mathbb{E}_{x, u, \epsilon \sim P_{id}} [((\theta_x^\top x + \theta_u^\top u + \epsilon) - \hat{\theta}_{x,ols}^\top x)^2] \quad (20)$$

$$= \mathbb{E}_{x \sim P_{id}} [(\theta_x^\top x - \hat{\theta}_{x,ols}^\top x)^2] + \mathbb{E}_{u \sim P_{id}} [\theta_u^\top u^2] + \mathbb{E}_{\epsilon \sim P_{id}} [\epsilon^2] \quad (21)$$

$$\geq \mathbb{E}_{u \sim P_{id}} [\theta_u^\top u^2] + \mathbb{E}_{\epsilon \sim P_{id}} [\epsilon^2] \quad (22)$$

$$= \sigma_u^2 + \sigma^2 \quad (23)$$

To get Equation 21, we expand the square, use linearity of expectation, and use the fact that x, u, ϵ are independent where u, ϵ are mean 0.

Upper bound risk of aux-inputs: On the other hand, we will show that if n is sufficiently large, the expected risk of the input model is less than $\sigma_u^2 + \sigma^2$.

First we show that we can write $y = \theta_x'^\top x + \theta_z'^\top z + \epsilon$ for some θ_x', θ_z' , that is y is a well-specified linear function of x and z plus some noise. Intuitively this is because y is a linear function of x, u and since C^* is invertible we can extract u from x, z . Formally, we assumed the true model is linear, that is, $y = \theta_x^\top x + \theta_u^\top u + \epsilon$. Since we have $z = A^* B^* x + C^* u$ where C^* is invertible, we can write $u = C^{*\top} (z - A^* B^* x)$. This gives us:

⁴ x and u are independent, with invertible covariance matrices, and $z = A^* B^* x + C^* u$ where C^* is full rank, so by block Gaussian elimination we can see that $[x, z]$ has invertible covariance matrix as well.

$$y = \theta_x^\top x + \theta_u^\top u + \epsilon \quad (24)$$

$$= \theta_x^\top x + \theta_u^\top C^{\star\top} (z - A^\star B^\star x) + \epsilon \quad (25)$$

$$= (\theta_x - B^{\star\top} A^{\star\top} (C^{\star\top})^{-1} \theta_u)^\top x + (C^{\star\top})^{-1} \theta_u^\top z + \epsilon \quad (26)$$

So setting $\theta'_x = \theta_x - B^{\star\top} A^{\star\top} (C^{\star\top})^{-1} \theta_u$ and $\theta'_z = C^{\star\top} (C^{\star\top})^{-1} \theta_u$, we get $y = \theta'^\top_x x + \theta'^\top_z z + \epsilon$.

As before, we note that the total mean squared error can be decomposed into the Bayes-opt error plus the excess error:

$$R_{\text{id}}(\hat{f}_{\text{in}}) = \mathbb{E}_{x, y, z \sim P_{\text{id}}} [(y - \hat{\theta}_{x, \text{in}}^\top x - \hat{\theta}_{z, \text{in}}^\top z)^2] \quad (27)$$

$$= \mathbb{E}_{x, z, \epsilon \sim P_{\text{id}}} [((\theta'^\top_x x + \theta'^\top_z z + \epsilon) - \hat{\theta}_{x, \text{in}}^\top x - \hat{\theta}_{z, \text{in}}^\top z)^2] \quad (28)$$

$$= \mathbb{E}_{x, z \sim P_{\text{id}}} [(\theta'^\top_x x + \theta'^\top_z z - \hat{\theta}_{x, \text{in}}^\top x - \hat{\theta}_{z, \text{in}}^\top z)^2] + \mathbb{E}_{\epsilon \sim P_{\text{id}}} [\epsilon^2] \quad (29)$$

$$= \mathbb{E}_{x, z \sim P_{\text{id}}} [(\theta'^\top_x x + \theta'^\top_z z - \hat{\theta}_{x, \text{in}}^\top x - \hat{\theta}_{z, \text{in}}^\top z)^2] + \sigma^2 \quad (30)$$

To get Equation 29, we expand the square, use linearity of expectation, and use the fact that x, z, ϵ are independent with $\mathbb{E}[\epsilon] = 0$. So it suffices to bound the excess error, $EE = \mathbb{E}_{x, z \sim P_{\text{id}}} [(\theta'^\top_x x + \theta'^\top_z z - \hat{\theta}_{x, \text{in}}^\top x - \hat{\theta}_{z, \text{in}}^\top z)^2]$.

To bound the excess error, we use Theorem 1 in Hsu et al. [20] where the inputs/covariates are $[x, z]$. $\mathbb{E}[[x, z][x, z]^\top]$ is invertible because $m = T$, C^\star is full rank, and P_x, P_u have density everywhere with density upper bounded so the variance in any direction is positive, and so the population covariance matrix is positive definite. This means $\mathbb{E}[[x, z][x, z]^\top]$ has min singular value lower bounded, and we also have that x, z are bounded random variables. Therefore, Condition 1 is satisfied for some finite ρ_0 . Condition 2 is satisfied since the noise ϵ is sub-Gaussian with mean 0 and variance parameter σ^2 . Condition 3 is satisfied with $b_0 = 0$, since we are working in the setting of well-specified linear regression.

To apply Theorem 1 [20], we first choose $t = \log \frac{3}{\delta}$ so that $1 - 3e^{-t} \geq 1 - \delta$, and so the statement of the Theorem holds with probability at least $1 - \delta$. Since our true model is linear (or as Remark 9 says that “the linear model is correct”), $\text{approx}(x) = 0$.

So as per remark 9 [20] Equation 11, for some constant c' , we have an upper bound on the excess error EE with probability at least $1 - \delta$:

$$EE \leq \frac{\sigma^2(d + 2\sqrt{dt} + 2t)}{n} + o(1/n) \quad (31)$$

Note that the notation in Hsu et al. [20] is different. The learned estimator in ordinary least squares regression is denoted by $\hat{\beta}_0$, the ground truth parameters by β , and the excess error is denoted by $\|\hat{\beta}_0 - \beta\|_\Sigma$. See section 2.1, 2.2 of Hsu et al. [20] for more details.

Since t is fixed, there exists some constant c (dependent on δ) such that for large enough N_1 if $n \geq N_1$:

$$EE \leq \sigma^2(cd/n) \quad (32)$$

Note that this is precisely Remark 10 [20]. Remark 10 says that $\|\hat{\beta}_0 - \bar{\beta}_0\|_\Sigma$ is within constant factors of $\sigma^2 d/n$ for large enough n . This is the variance term, but the bias term is 0 since the linear model is well-specified so $\text{approx}(x) = 0$. As in Proposition 2 [20] the total excess error is bounded by 2 times the sum of the bias and variance term, which gives us the same result.

Putting this (Equation 32) back into Equation 30, we get that with probability at least $1 - \delta$:

$$R_{\text{id}}(\hat{f}_{\text{in}}) \leq \sigma^2(1 + cd/n) \quad (33)$$

Since $\sigma_u^2 > 0$, we have $\sigma^2 < \sigma_u^2 + \sigma^2$. Then for some N and for all $n \geq N$, we have:

$$R_{\text{id}}(\hat{f}_{\text{in}}) < \sigma_u^2 + \sigma^2 \leq R_{\text{id}}(\hat{f}_{\text{bs}}) \quad (34)$$

In particular, we can choose $N = \max(N_0, c \frac{\sigma^2}{\sigma_u^2} d + 1)$. Which completes the proof. \square

E.3 Auxiliary inputs can hurt out-of-distribution

Restatement of Example 1. *There exists a problem setting \mathcal{P}, P_ϵ , such that for every n , there is some test distribution P'_x, P'_u with:*

$$\mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{in}})] > \mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{bs}})] \quad (35)$$

Proof. We will have $x \in \mathbb{R}$ (so $d = 1$), $w = x$, and $u, z \in \mathbb{R}^2$. We set $z_1 = u_1 + w$ and $z_2 = u_2$, in other words we choose $A^* = [1, 0]$ and $C^* = I_2$ is the identity matrix. We set $y = x + u_1 + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$, so y is a function of x and u_1 but not u_2 . In other words we choose $\theta_w = 1$ and $\theta_u = (1, 0)$. P_x will be Uniform $[-1, 1]$, and P_u will be uniform in the unit ball in \mathbb{R}^2 .

Let $X_Z = [X; Z]$, which denotes appending X and Z by columns so $X_Z \in \mathbb{R}^{n \times 3}$ with $n \geq 3$. Since P_x and P_u have density, X_Z has rank 3 almost surely. This means that $X_Z^\top X_Z$ is invertible (and positive semi-definite) almost surely. Since P_x and P_u are bounded, the maximum eigenvalue τ'_{max} of $X_Z^\top X_Z$ is bounded above. The minimum eigenvalue τ'_{min} of $(X_Z^\top X_Z)^{-1}$ is precisely $1/\tau'_{max}$ and is therefore positive and bounded below by some $c > 0$ almost surely.

We will define P'_x and P'_u soon. For now, consider a new test example $x' \sim P'_x, u' \sim P'_u$ with $z' = [x', 0] + u'$ and $y' = x' + u'_1 + \epsilon'$ with $\epsilon' \sim N(0, \sigma^2)$ and $\mathbb{E}[u'] = 0$. For the input model we have:

$$\mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{in}})] = \mathbb{E}[(y' - (\hat{\theta}_{x, \text{in}}^\top x' + \hat{\theta}_{z, \text{in}}^\top z'))^2] \quad (36)$$

$$= \sigma^2(1 + \mathbb{E}[(x', z')^\top (X_Z^\top X_Z)^{-1} (x', z')]) \quad (37)$$

$$\geq \sigma^2(1 + \mathbb{E}[\tau'_{min} \|(x', z')\|_2^2]) \quad (38)$$

$$\geq \sigma^2(1 + c \mathbb{E}[\|(x', z')\|_2^2]) \quad (39)$$

$$\geq \sigma^2(1 + c \mathbb{E}[z_2'^2]) \quad (40)$$

$$= \sigma^2(1 + c \mathbb{E}[u_2'^2]) \quad (41)$$

Notice that this lower bound is a function of $\mathbb{E}[u_2'^2]$ which we will make very large.

On the other hand, letting $\sigma_u'^2 = \mathbb{E}_{u' \sim P'_u}[(\theta_u^\top u')^2] = \mathbb{E}_{u' \sim P'_u}[u_1'^2]$, for the baseline model we have:

$$\mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{bs}})] = \mathbb{E}[(y' - \hat{\theta}_{x, \text{ols}}^\top x')^2] \quad (42)$$

$$= (\sigma^2 + \sigma_u'^2) + (\sigma^2 + \sigma_u'^2) \mathbb{E}[x'^\top (X^\top X)^{-1} x'] \quad (43)$$

So the risk depends on x' and $\mathbb{E}[u_1'^2]$ but not $\mathbb{E}[u_2'^2]$.

So we choose $P'_x = \text{Uniform}(-1, 1)$. For P'_u , we sample the components independently, with $u_1' \sim \text{Uniform}(-1, 1)$, and $u_2' \sim \text{Uniform}(-R, R)$. By choosing R large enough, we can make the lower bound for the input model arbitrarily large without impacting the risk of the baseline model which gives us:

$$\mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{in}})] > \mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{bs}})] \quad (44)$$

□

E.4 Pre-training improves risk under arbitrary covariate shift

Restatement of Theorem 3. For all problem settings \mathcal{P} , P_e , and for all test distributions P'_x and P'_u :

$$\mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{out}})] \leq \mathbb{E}[R_{\text{ood}}(\hat{f}_{\text{bs}})] \quad (45)$$

First we show that pre-training (training a low-rank linear map from x to z) recovers the unobserved features w . We will then show that learning a regression map from w to y is better in all directions than learning a regression map from x to y .

Our first lemma shows that we can recover the map from x to w up to identifiability (we will learn the rowspace of the true linear map from x to w).

Lemma 1. For a pair (x, z) , let $z = A^* B^* x + \xi$ where $A^* \in \mathbb{R}^{T \times k}$ and $B^* \in \mathbb{R}^{k \times d}$ are the true parameters with $T, d \geq k$ and $\xi \in \mathbb{R}^T$ is mean-zero noise with bounded variance in each coordinate. Assume that A^*, B^* are both rank k . Suppose that $\mathbb{E}[xx^\top]$ is invertible. Let \hat{A}, \hat{B} be minimizers of the population risk $\mathbb{E}[\|\hat{A}\hat{B}x - z\|^2]$ of the multiple-output regression problem. Then $\text{span}\{B^*_1, \dots, B^*_k\} = \text{span}\{\hat{B}_1, \dots, \hat{B}_k\}$ where B^*_i, \hat{B}_i are the i -th rows of their respective matrices.

Proof. We first consider solving for the product of the weights $\hat{A}\hat{B}$. The population risk can be decomposed into the risks of the T coordinates of the output:

$$L(C) = \mathbb{E}[\|Cx - z\|^2] \quad (46)$$

$$= \mathbb{E}[\|Cx - A^*B^*x - \xi\|^2] \quad (47)$$

$$= \mathbb{E}[\|(C - A^*B^*)x\|^2] + \mathbb{E}[\|\xi\|^2] \quad (48)$$

$$= \sum_{i=1}^T \mathbb{E}[(C_i - (A^*B^*)_i)x]^2 + \mathbb{E}[\xi^2] \quad (49)$$

$$= \sum_{i=1}^T L_i(C) \quad (50)$$

where $\mathbb{E}[\|\xi\|^2]$ is the error of the Bayes optimal estimator, $C_i, (A^*B^*)_i$ denote the i -th row of the respective matrices, and the i -th risk is $L_i(C) := \mathbb{E}[(C_i - (A^*B^*)_i)x]^2 + \mathbb{E}[\xi^2]$. From this decomposition, we see that for every i ,

$$(A^*B^*)_i = \underset{C_i}{\operatorname{argmin}} L_i(C_i) \implies A^*B^* = \underset{C}{\operatorname{argmin}} L(C) \quad (51)$$

such that A^*B^* is the unique minimizer of the population risk for the multiple-output regression problem. The minimizer is unique since the minimizer for each subproblem is unique because $\mathbb{E}[xx^\top]$ is invertible. Since the variance of ξ is bounded in each coordinate, with infinite data we achieve the same minimizer of the multiple-output regression problem. Thus we have

$$\hat{A}\hat{B} = \underset{C}{\operatorname{argmin}} L(C) = A^*B^* \quad (52)$$

so that the product of the learned parameters and the true parameters are equal.

From this, we see that \hat{A}, \hat{B} must be rank k , or else $\hat{A}\hat{B} \neq A^*B^*$. Since \hat{A} is rank k , it has a left inverse $\hat{A}_{\text{left}}^{-1}$ such that

$$\hat{B} = Q B^* \quad (53)$$

where $Q = \hat{A}_{\text{left}}^{-1} A^* \in \mathbb{R}^{k \times k}$ and $\hat{A}_{\text{left}}^{-1} = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top$. Since $\hat{A}_{\text{left}}^{-1}$ is rank k and A^* is rank k , we have that the rank of the product is $\operatorname{rank}(Q) \leq k$. In the other direction, by Sylvester's rank inequality, $\operatorname{rank}(Q) \geq \operatorname{rank}(\hat{A}_{\text{left}}^{-1}) + \operatorname{rank}(A^*) - k = k$, which implies that Q is full rank (rank k). This implies the result. \square

Before proving the theorem, it will be useful to state a version of the Woodbury inversion lemma.

Lemma 2 (Woodbury). *For invertible $A \in \mathbb{R}^{d \times d}$, invertible $C \in \mathbb{R}^{k \times k}$, and $D \in \mathbb{R}^{d \times k}$,*

$$(A + DC D^\top)^{-1} = A^{-1} - A^{-1} D (C^{-1} + D^\top A^{-1} D)^{-1} D^\top A^{-1}. \quad (54)$$

Our next lemma shows that for any fixed training examples X and arbitrary test example x' , the aux-outputs model will have better expected risk than the baseline where the expectation is taken over the training labels $Y | X$.

Lemma 3. *In the linear setting, fix data matrix X and consider arbitrary test example x' . Let $\theta^* = B^{*\top} \theta_w$ be the optimal (ground truth) linear map from x to y . The expected excess risk of the aux-outputs model $\hat{B}^\top \hat{\theta}_{w, \text{out}}$ is better than for the baseline $\hat{\theta}_{x, \text{ols}}$, where the expectation is taken over the training targets $Y \sim P_{Y|X}$ (Y shows up implicitly because the estimators $\hat{\theta}_{w, \text{out}}$ and $\hat{\theta}_{x, \text{ols}}$ depend on Y):*

$$\mathbb{E}[(\hat{\theta}_{w, \text{out}}^\top \hat{B} x' - \theta^{*\top} x')^2] \leq \mathbb{E}[(\hat{\theta}_{x, \text{ols}}^\top x' - \theta^{*\top} x')^2] \quad (55)$$

Proof. Let $\epsilon_{\text{all}} = Y - X\theta^*$ be the training noise. From standard calculations, the instance-wise risk of $\hat{\theta}_{x, \text{ols}}$ for any x is

$$\mathbb{E}[(\hat{\theta}_{x, \text{ols}}^\top x' - \theta^{*\top} x')^2] = \mathbb{E}[(X^\top X)^{-1} X^\top Y]^\top x' - \theta^{*\top} x')^2] \quad (56)$$

$$= \mathbb{E}[(\theta^* + (X^\top X)^{-1} X^\top \epsilon_{\text{all}})^\top x' - \theta^{*\top} x')^2] \quad (57)$$

$$= \mathbb{E}[(X^\top X)^{-1} X^\top \epsilon_{\text{all}}]^\top x')^2] \quad (58)$$

$$= (\sigma^2 + \sigma_u^2) x'^\top (X^\top X)^{-1} x' \quad (59)$$

By Lemma 1, $\hat{B} = QB$ for some full rank Q . Thus, learning $\hat{\theta}_{w, \text{out}}$ is a regression problem with independent mean-zero noise and we can apply the same calculations for the instance-wise risk of $\hat{B}^\top \hat{\theta}_{w, \text{out}}$.

$$\mathbb{E}[(\hat{\theta}_{w, \text{out}}^\top \hat{B} x' - \theta^{*\top} x')^2] = (\sigma^2 + \sigma_u^2) x'^\top \hat{B}^\top (\hat{B} X^\top X \hat{B}^\top)^{-1} \hat{B} x'. \quad (60)$$

We show that the difference between the inner matrices is positive semi-definite, which implies the result. In particular, we show that

$$(X^\top X)^{-1} - \hat{B}^\top (\hat{B} X^\top X \hat{B}^\top)^{-1} \hat{B} \succcurlyeq 0. \quad (61)$$

Multiplying the Woodbury inversion lemma on the left and right by A , we have the identity

$$A(A + DCD^\top)^{-1}A = A - D(C^{-1} + D^\top A^{-1}D)^{-1}D^\top. \quad (62)$$

Letting $A = (X^\top X)^{-1}$, $C = \alpha I$, and $D = B^\top$, we have

$$(X^\top X)^{-1} - \hat{B}^\top (\hat{B} X^\top X \hat{B}^\top)^{-1} \hat{B} = A - D(D^\top A^{-1}D)^{-1}D^\top \quad (63)$$

$$= \lim_{\alpha \rightarrow \infty} A - D\left(\frac{1}{\alpha}I + D^\top A^{-1}D\right)^{-1}D^\top \quad (64)$$

$$= \lim_{\alpha \rightarrow \infty} A(A + \alpha DD^\top)^{-1}A. \quad (65)$$

where the last step applies Equation (62). The first limit converges by continuity of the inverse, and the second limit converges since the elements of the sequence are identical to the first. For any $\alpha > 0$, $A + \alpha DD^\top$ is a sum of PSD matrices and is thus PSD. Since A is symmetric and PD, each element of the limit sequence is PSD. Since the space of PSD matrices is closed, we have that the original difference of matrices is PSD. This implies the result. \square

Proof of Theorem 3. Fix training examples X and test example x' but let the train labels $Y \sim P_{Y|X}$ and and test label $y' \sim P'_{y'|x'}$ be random. In particular, let $\sigma_u'^2 = \mathbb{E}[(\theta_u^\top u')^2]$ where $u' \sim P'_u$, with $\mathbb{E}[u'] = 0$. Then for the baseline OLS estimator, we have:

$$\mathbb{E}[(y' - \hat{\theta}_{x,ols}^\top x')^2] = \sigma_u'^2 + \sigma^2 + \mathbb{E}[(\hat{\theta}_{x,ols}^\top x' - \theta^{\star\top} x')^2] \quad (66)$$

For the aux-outputs model, we have:

$$\mathbb{E}[(y' - \hat{\theta}_{w,out}^\top \hat{B}x')^2] = \sigma_u'^2 + \sigma^2 + \mathbb{E}[(\hat{\theta}_{w,out}^\top \hat{B}x' - \theta^{\star\top} x')^2] \quad (67)$$

So applying Lemma 3, we get that the risk for the aux-outputs model is better than for the baseline (the lemma showed it for the excess risk):

$$\mathbb{E}[(y' - \hat{\theta}_{w,out}^\top \hat{B}x')^2] \leq \mathbb{E}[(y' - \hat{\theta}_{x,ols}^\top x')^2] \quad (68)$$

Since this is true for all X and x' , it holds when we take the expectation over the training examples X from P_x and the test example x' from $P'_{x'}$ which gives us the desired result. \square

E.5 In-N-Out improves risk under arbitrary covariate shift

Restatement of Theorem 4. *In the linear setting, for all problem settings \mathcal{P} with $\sigma_u^2 > 0$, test distributions P'_x, P'_u , and $\delta > 0$, there exists $a, b > 0$ such that for all P_e , with probability at least $1 - \delta$ over the training examples and test example $x' \sim P'_{x'}$, the ratio of the excess risks (for all σ^2 small enough that $a - b\sigma^2 > 0$) is:*

$$\frac{R_{in-out}^{ood} - R^*}{R_{out}^{ood} - R^*} \leq \frac{\sigma^2}{a - b\sigma^2} \quad (69)$$

Here $R^* = \min_{f^*} R_{ood}(f^*)$ is the minimum possible (Bayes-optimal) OOD risk, $R_{in-out}^{ood} = \mathbb{E}_{y' \sim P'_{y'|x'}}[\ell(\hat{g}(\hat{h}_{out}(x')), y')]$ is the risk of the In-N-Out model on test example x' , and $R_{out}^{ood} = \mathbb{E}_{y' \sim P'_{y'|x'}}[\ell(\hat{g}_{out}^y(\hat{h}_{out}(x')), y')]$ is the risk of the aux-outputs model on test example x' .

We first show a key lemma that lets us bound the min singular values of a random matrix, which will let us upper bound the risk of the In-N-Out estimator and lower bound the risk of the pre-training estimator.

Definition 1. *As usual, the min singular value $\tau_{min}(W)$ of a rectangular matrix $W \in \mathbb{R}^{n \times k}$ where $n \geq k$ refers to the k -th largest singular value (the remaining $n - k$ singular values are all 0), or in other words:*

$$\tau_{min}(W) = \min_{\|\nu\|_2=1} \|W\nu\|_2 \quad (70)$$

Lemma 4. *Let P_w and P_u be independent distributions on \mathbb{R}^k and \mathbb{R}^m respectively. Suppose they are absolutely continuous with respect to the standard Lebesgue measure on \mathbb{R}^k and \mathbb{R}^m respectively (e.g. this is true if they have density everywhere with density upper bounded). Let $W \in \mathbb{R}^{n \times k}$ where each row W_i is sampled independently $W_i \sim P_w$. Let $U \in \mathbb{R}^{n \times m}$ where each row U_i is sampled independently $U_i \sim P_u$. Suppose $n \geq k + m$. For all δ , there exists $c(\delta) > 0$ such that with probability at least $1 - \delta$, the minimum singular values τ_{min} are lower bounded by $c(\delta)$: $\tau_{min}(W) > c(\delta)$ and $\tau_{min}([W; U]) > c(\delta)$.*

Proof. We note that the matrices W and U are rectangular, e.g. $W \in \mathbb{R}^{n \times k}$ where $n \geq k$. We will prove the lemma for W first, and the extension to $[W; U]$ will follow.

Note that removing the last $n - k$ rows of W cannot increase its min singular value since that corresponds to projecting the vector $W\nu$ and projection never increases the Euclidean norm. So WLOG we suppose W only consists of its first k rows and so $W \in \mathbb{R}^{k \times k}$.

Now, consider any row W_i . We will use a volume argument to show that with probability at least $1 - \frac{\delta}{d}$, this row W_i has a non-trivial component perpendicular to all the other rows. Since all rows are independently and identically sampled, without loss of generality suppose $i = 1$. Fix the other rows w_2, \dots, w_k , since w_1 is independent of these other rows, the conditional distribution of w_1 is the same as the marginal of w_1 . w_2, \dots, w_k form a $k - 1$ dimensional subspace S in \mathbb{R}^k . Letting $d(w, S)$ denote the Euclidean distance of a vector w from the subspace S , define the event $S_\lambda = \{w_1 : d(w_1, S) \leq \lambda\}$. Since P_w is absolutely continuous, $P(S_\lambda) \rightarrow 0$ as $\lambda \rightarrow 0$, so for some small $c(\delta) > 0$, $P(S_{c(\delta)}) < \frac{\delta}{d}$. So with probability at least $1 - \delta/d$, $d(w_1, S) > c(\delta)$.

By union bound, with probability at least $1 - \delta$, this is true for every row w_i —condition on the event that this is true. By representing each row vector as the sum of the component perpendicular to the subspace of the other vectors, and a component along the subspace, and applying Pythagoras theorem and expanding we get:

$$\min_{\|\nu\|_2=1} \|W\nu\| = \min_{\|\nu\|_2=1} \|\nu^\top W\| \geq c(\delta) \quad (71)$$

Which completes the proof for $\tau_{\min}(W)$.

For $[W; U]$, we note that P_x and P_u are independent, and the product measure is absolutely continuous. Since each row of $[W; U]$ is identically and independently sampled just like with W , we can apply the exact same argument as above (though for a different constant $c(\delta)$, we take the min of these two as our $c(\delta)$ in the lemma statement). \square

Recall that the In-n-Out estimator was obtained by fitting a model from w, z to y , and then using that to produce pseudolabels on (infinite) unlabeled data, and then self-training a model from w to y on these pseudolabels. For the linear setting, we defined the In-N-Out estimator $\hat{\theta}_w$ in Equation 17. Our next Lemma gives an alternate closed form of the In-N-Out estimator in terms of the representation matrix $W = X\hat{B}$ and the latent matrix U .

Lemma 5. *In the linear setting, letting $W = X\hat{B}^\top$ we can write the In-n-Out estimator in closed form as:*

$$\hat{\theta}_w = [I_{k \times k}; 0_{k \times T}] \left(\begin{pmatrix} W^\top \\ U^\top \end{pmatrix} (W; U) \right)^{-1} \begin{pmatrix} W^\top \\ U^\top \end{pmatrix} Y \quad (72)$$

Proof. We recall the definition of the In-N-Out estimator, where we first train a classifier from W, Z to Y .

$$\hat{\gamma}_w, \hat{\gamma}_z = \underset{\gamma_w, \gamma_z}{\operatorname{argmin}} \|(Y - (W\gamma_w + Z\gamma_z))^2\| \quad (73)$$

Denote the minimum value of Equation 73 by p^* . Note that $\hat{\gamma}_w, \hat{\gamma}_z$ may not be unique, and we pick any solution to the argmin (although our proof will reveal that the resulting $\hat{\theta}_w$ is in fact unique). We then use this to produce pseudolabels and self-train, on infinite data, which gives us the In-N-Out estimator:

$$\hat{\theta}_w = \hat{\gamma}_w + A^\top \hat{\gamma}_z \quad (74)$$

We will now consider the following alternative estimator:

$$\hat{\theta}'_w, \hat{\theta}'_u = \underset{\hat{\theta}'_w, \hat{\theta}'_u}{\operatorname{argmin}} \|(Y - (W\hat{\theta}'_w + U\hat{\theta}'_u))^2\| \quad (75)$$

Denote the minimum value of Equation 75 by q^* . We claim that $\hat{\theta}'_w = \hat{\theta}_w$.

We will show that the In-N-Out estimator $\hat{\theta}_w$ minimizes the alternative minimization problem in Equation 75 by showing that $p^* = q^*$. We will then show that the solution to Equation 75 is unique, which implies that $\hat{\theta}'_w = \hat{\theta}_w$.

We note that $C^{*\top} \in \mathbb{R}^{m \times T}$ where $T \geq m$ is full-rank, so there exists a right-inverse C' with $C^{*\top} C' = I_{m \times m}$. Since $Z = W A^{*\top} + U C^{*\top}$, this gives us: $U = (Z - W A^{*\top}) C' = Z C' + W (-A^{*\top} C')$.

So this means that a solution to the alternative problem in Equation 75 can be converted into a solution for the original in Equation 73 with the same function value:

$$\min_{\hat{\theta}'_w, \hat{\theta}'_u} \|(Y - (W\hat{\theta}'_w + U\hat{\theta}'_u))^2\| \quad (76)$$

$$= \min_{\hat{\theta}'_w, \hat{\theta}'_u} \|(Y - (W\hat{\theta}'_w + (ZC' + W(-A^{*\top}C'))\hat{\theta}'_u))^2\| \quad (77)$$

$$= \min_{\hat{\theta}'_w, \hat{\theta}'_u} \|(Y - (W(\hat{\theta}'_w - A^{*\top}C'\hat{\theta}'_u) + Z(C'\hat{\theta}'_u)))^2\| \quad (78)$$

This implies that $p^* \leq q^*$.

We now show that a solution to the original problem in Equation 73 can be converted into a solution for the alternative in Equation 75 with the same function value:

$$\min_{\hat{\gamma}_w, \hat{\gamma}_z} \|(Y - (W\hat{\gamma}_w + Z\hat{\gamma}_z))^2\| \quad (79)$$

$$= \min_{\hat{\gamma}_w, \hat{\gamma}_z} \|(Y - (W\hat{\gamma}_w + (WA^{*\top} + UC^{*\top})\hat{\gamma}_z))^2\| \quad (80)$$

$$= \min_{\hat{\gamma}_w, \hat{\gamma}_z} \|(Y - (W(\hat{\gamma}_w + A^{*\top}\hat{\gamma}_z) + U(C^{*\top}\hat{\gamma}_z)))^2\| \quad (81)$$

This implies that $q^* \leq p^*$, and we showed before that $p^* \leq q^*$ so $p^* = q^*$. But since $\hat{\gamma}_w, \hat{\gamma}_z$ minimizes the original minimizer in Equation 73, $\hat{\gamma}_w + A^{*\top}\hat{\gamma}_z, C^{*\top}\hat{\gamma}_z$ minimize the alternative problem in Equation 75, where $\hat{\theta}'_w = \hat{\gamma}_w + A^{*\top}\hat{\gamma}_z$.

Since $[W; U]$ is full rank, the solution $\hat{\theta}'_w, \hat{\theta}'_u$ to the alternative estimator Equation 75 is unique. So this means that $\hat{\theta}'_w = \hat{\theta}_w$.

We have shown that $\hat{\theta}'_w = \hat{\theta}_w$ —this completes the proof because solving Equation 75 for $\hat{\theta}'_w$ gives us the closed form in Equation 72.

□

Next we show a technical lemma that says that if a random vector $u \in \mathbb{R}^n$ has bounded density everywhere, then for any v with high probability the dot product $(u^\top v)^2$ cannot be too small relative to $\|v\|_2^2$.

Lemma 6. *Suppose a random vector $u \in \mathbb{R}^n$ has density everywhere, with bounded density. For every δ , there exists some $c(\delta)$ such that for all v , with probability at least $1 - \delta$ over u , $(u^\top v)^2 \geq c(\delta)\|v\|_2^2$.*

Proof. First, we choose some B_0 such that $P(\|u\|_2 \geq B_0) \leq \delta/2$, such a B_0 exists for every probability measure.

Suppose that the density is upper bounded by B_1 . Let the area of the $n - 1$ dimensional sphere with radius B_0 be A_0 . Consider any $n - 1$ dimensional subspace S , and let $S_\epsilon = \{u' : d(u', S) \leq \epsilon\}$ where $d(u', S)$ denotes the Euclidean distance from u' to S . $P(u \in S) \leq A_0 B_1 \epsilon + \delta/2$ for all S . By choosing sufficiently small $\epsilon > 0$, we can ensure that $P(u \in S) \leq \delta$ for all S .

Now consider arbitrary v and let $S(v)$ be the $n - 1$ -dimensional subspace perpendicular to v . We have $P(u \in S(v)_\epsilon) \leq \delta$. But this means that $(u^\top v)^2 \geq \epsilon^2\|v\|_2^2$ with probability at least $1 - \delta$, which completes the proof. □

By definition of our linear multi-task model, we recall that $y = \theta_w^\top w + \theta_u^\top u + \epsilon$, where $w = B^*x$. We do not have access to B^* , but we assumed that B^* is full rank. We learned \hat{B} which has the same rowspace as B^* (Lemma 1). This means that for some $\hat{\theta}'_w$, we have $y = \hat{\theta}'_w^\top \hat{w} + \theta_u^\top u + \epsilon$ where $\hat{w} = \hat{B}x$. To simplify notation and avoid using θ'_w and \hat{w} everywhere, we suppose WLOG that $\hat{B} = B^*$ (but formally, we can just replace all the occurrences of θ_w by θ'_w and w by \hat{w}).

Our next lemma lower bounds the test error of the pre-training model.

Lemma 7. *In the linear setting, for all problem settings \mathcal{P} with $\sigma_u^2 > 0$, for all δ , there exists some $a, b > 0$ such that with probability at least $1 - \delta$ over the training examples and test example $x' \sim P'_x$ the risk of the aux-outputs model is lower bounded:*

$$R_{out}^{ood} - R^* > a - b\sigma^2 \quad (82)$$

Proof. Recall that $R_{\text{out}}^{\text{ood}} = \mathbb{E}_{y' \sim P'_{y'|x'}} [l(\hat{f}_{\text{out}}(\hat{h}_{\text{out}}(x')), y')]$. Let $\sigma_u'^2 = \mathbb{E}_{u' \sim P'_u} [(\theta_u^\top u')^2]$. We have $R^* = \sigma^2 + \sigma_u'^2$. Let $W = XB^{\star\top}$ be the feature matrix, where $W \in \mathbb{R}^{n \times k}$.

Letting $w' = B^* x'$, for the aux-outputs model, we have:

$$\mathbb{E}_{y' \sim P'_{y'|x'}} [l(\hat{f}_{\text{out}}(\hat{h}_{\text{out}}(x')), y')] \quad (83)$$

$$= \mathbb{E}_{y' \sim P'_{y'|x'}} [(y' - \hat{\theta}_{w, \text{out}}^\top w')^2] \quad (84)$$

$$= (\sigma^2 + \sigma_u'^2) + (\theta_w^\top w' - \hat{\theta}_{w, \text{out}}^\top w')^2 \quad (85)$$

$$= R^* + (\theta_w^\top w' - \hat{\theta}_{w, \text{out}}^\top w')^2 \quad (86)$$

Let $\epsilon = Y - (W\theta_w + U\theta_u)$ be the noise of Y for the training examples, which is a random vector with $\epsilon \in \mathbb{R}^n$. We can now write:

$$(\theta_w^\top w' - \hat{\theta}_{w, \text{out}}^\top w')^2 = ((\epsilon + U\theta_u)^\top W(W^\top W)^{-1} w')^2 \quad (87)$$

By assumption, $W^\top W$ is invertible (almost surely). With probability at least $1 - \delta/10$ all entries of $W^\top W$ are upper bounded and we condition on this. So $(W^\top W)^{-1}$ has min singular value bounded below. By Lemma 4, W has min singular value that is bounded below with probability at least $1 - \delta/10$, we condition on this being true. So let $\nu = W(W^\top W)^{-1} w'$, so for some $c_0 > 0$, we have: $\|\nu\|_2 \geq c_0 \|w'\|_2$.

In terms of ν , we can write Equation 87 as:

$$(\theta_w^\top w' - \hat{\theta}_{w, \text{out}}^\top w')^2 = ((\epsilon + U\theta_u)^\top \nu)^2 \quad (88)$$

$$= (\epsilon^\top \nu)^2 + ((U\theta_u)^\top \nu)^2 + 2(\epsilon^\top \nu)((U\theta_u)^\top \nu) \quad (89)$$

$$\geq ((U\theta_u)^\top \nu)^2 + 2(\epsilon^\top \nu)((U\theta_u)^\top \nu) \quad (90)$$

$$\geq ((U\theta_u)^\top \nu)^2 - 2|\epsilon^\top \nu| \|(U\theta_u)^\top\|_2 \|\nu\|_2 \quad (91)$$

We can find b_0 such that with at least probability $1 - \delta/10$, $\|(U\theta_u)^\top\|_2 \leq b_0$, condition on this. We note that $\epsilon^\top \nu$ has variance $\sigma^2 \|\nu\|_2$ so by Chebyshev for some b_1 with probability at least $1 - \delta$, $|\epsilon^\top \nu| \leq b_1 \sigma^2 \|\nu\|_2$, condition on this. So we can now bound Equation 91 and get:

$$(\theta_w^\top w' - \hat{\theta}_{w, \text{out}}^\top w')^2 \geq ((U\theta_u)^\top \nu)^2 - 2b_0 b_1 \sigma^2 \|\nu\|_2^2 \quad (92)$$

Now we apply Lemma 6, where we use the fact that $\sigma_u'^2 > 0$. So given $\delta/10$, there exists some c_1 such that for every ν with probability at least $1 - \delta/10$, $((U\theta_u)^\top \nu)^2 \geq c_1 \|\nu\|_2^2$, giving us:

$$(\theta_w^\top w' - \hat{\theta}_{w, \text{out}}^\top w')^2 \geq (c_1 - 2b_0 b_1 \sigma^2) \|\nu\|_2^2 \quad (93)$$

Since w' has bounded density everywhere, it is non-atomic and we get that there is some $c_2 > 0$ such that with probability at least $1 - \delta/10$, $\|w'\|_2^2 \geq c_2^2$. But then $\|\nu\|_2^2 \geq c_0 c_2$, which gives us for some a, b :

$$(\theta_w^\top w' - \hat{\theta}_{w, \text{out}}^\top w')^2 \geq (c_1 - 2b_0 b_1 \sigma^2) c_0 c_2 \geq a - b\sigma^2 \quad (94)$$

Combining this with Equation 86, we get with probability at least $1 - \delta$:

$$\mathbb{E}_{y' \sim P'_{y'|x'}} [l(\hat{f}_{\text{out}}(\hat{h}_{\text{out}}(x')), y')] - R^* > a - b\sigma^2 \quad (95)$$

Which was what we wanted. \square

Lemma 8. *In the linear setting, for all problem settings \mathcal{P} , for all δ , there exists some $c > 0$ such that with probability at least $1 - \delta$ over the training examples and test example $x' \sim P'_x$ the risk of the In-N-Out model is upper bounded:*

$$R_{\text{in-out}}^{\text{ood}} - R^* < c\sigma^2 \quad (96)$$

Proof. Recall that $R_{\text{in-out}}^{\text{ood}} = \mathbb{E}_{y' \sim P'_{y'|x'}} [l(\hat{f}_{\text{out}}(\hat{h}_{\text{out}}(x')), y')]$. Let $\sigma_u'^2 = \mathbb{E}_{u' \sim P'_u} [(\theta_u^\top u')^2]$. We have $R^* = \sigma^2 + \sigma_u'^2$. As before, let $W = XB^{\star\top}$ be the feature matrix, where $W \in \mathbb{R}^{n \times k}$.

Let $W_U = [W; U]$ which denotes concatenating the matrices by column, so that $W_U \in \mathbb{R}^{n \times (k+m)}$. By Lemma 4, $W_U^\top W_U$ has min singular value that is bounded below by c_0 with probability at least $1 - \delta/10$, we condition on this being true. Now, as for the aux-outputs model, letting $w' = B^* x'$, we have:

$$\mathbb{E}_{y' \sim P'_{y'|w'}} [(y' - \hat{\theta}_w^\top w')^2] = R^* + (\theta_w^\top w' - \hat{\theta}_w^\top w')^2 \quad (97)$$

For the second term on the RHS: Let $R = [I_{k \times k}; 0_{k \times m}]$. Let $\epsilon = Y - (W\theta_w + U\theta_u)$ be the noise of Y for the training examples, which is a random vector with $\epsilon \in \mathbb{R}^n$. We can now write:

$$(\theta_w^\top w' - \hat{\theta}_w^\top w')^2 = (w'^\top R(W_U^\top W_U)^{-1} W_U^\top \epsilon)^2 \quad (98)$$

$\|w'\|_2$ is bounded above by some constant B_1 with probability at least $1 - \delta/10$ which we condition on. Now taking the expectation over w' and ϵ , using the fact that R preserves the norm of a vector we can write:

$$\mathbb{E}_{w', \epsilon} [(w'^\top R(W_U^\top W_U)^{-1} W_U^\top \epsilon)^2] \quad (99)$$

$$= \sigma^2 \mathbb{E}_{w', \epsilon} [(w'^\top R[W_U^\top W_U]^{-1} R^\top w')] \quad (100)$$

$$\leq \frac{\sigma^2}{c_0^2} \mathbb{E}_{w'} [\|w'\|_2^2] \quad (101)$$

$$\leq \frac{B_1^2 \sigma^2}{c_0^2} \quad (102)$$

Then, by Markov's inequality, with probability at least $1 - \delta/10$ we can upper bound this by $\frac{10B_1^2 \sigma^2}{\delta c_0^2}$. In total, that gives us that for some c , with probability at least $1 - \delta$:

$$\mathbb{E}_{y' \sim P'_{y'|x'}} [l(\hat{f}_{\text{out}}(\hat{h}_{\text{out}}(x')), y')] - R^* < c\sigma^2 \quad (103)$$

□

The proof of Theorem 4 simply combines Lemma 7 and Lemma 8.

Proof of Theorem 4. For some a, b, c , with probability at least $1 - \delta$, we have for the aux-outputs model:

$$R_{\text{out}}^{\text{ood}} - R^* > a - b\sigma^2 \quad (104)$$

And for the In-N-Out model:

$$R_{\text{in-out}}^{\text{ood}} - R^* < c\sigma^2 \quad (105)$$

Taking ratios and dividing by suitable constants we get the desired result. □

F Experimental details

Data splits. In all experiments we first split off the OOD data, then randomly split the rest into training, validation, and in-distribution test. We use a portion of the training and OOD set as in-distribution and OOD unlabeled data respectively. The rest of the OOD set is held-out. We run 5 trials, where we regenerate the training/unlabeled split for each trial (keeping held-out splits fixed). We use a reduced number of labeled examples from each dataset (1%, 5%, 10% of labeled examples for CelebA, Cropland, and Landcover respectively), with the rest as unlabeled.

F.1 Cropland

All models reported in Table 1 were trained using the Adam optimizer with learning rate 0.001, a batch size of 256, and 100 epochs unless otherwise specified. Our dataset consists of about 7k labeled examples, 170k unlabeled examples (with 130k in-distribution examples), 7.5k examples each for validation and in-distribution test, and 4260 OOD test examples (the specification of OOD points is described in further detail below). Results are reported over 5 trials, and $\lambda \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ was chosen using the validation set.

Problem Motivation. Developing machine learning models trained on remote sensing data is currently a popular line of work for practical problems such as typhoon rainfall estimation, monitoring reservoir water quality, and soil moisture estimation [29, 32, 1]. Models that could use remote sensing data to accurately forecast crop yields or estimate the density of regions dedicated to growing crops would be invaluable in important tasks like estimating a developing nation's food security [30].

OOD Split. In remote sensing problems it is often the case that certain regions lack labeled data (e.g., due to a lack of human power to gather the labels on site), so extrapolation to these unlabeled regions is necessary. To simulate this data regime, we use the provided (lat, lon) pairs of each data point to split the dataset into labeled (in-distribution) and unlabeled (out-of-distribution) portions. Specifically, we take all points lying in Iowa, Missouri, and Illinois as our ID points and use all points within Indiana and Kentucky as our OOD set.

Shape of auxiliary info. To account for the discrepancy in shapes of the two sources of auxiliary information (latitude and longitude are two scalar measurements while the 3 vegetation bands form a $3 \times 50 \times 50$ tensor), we create latitude and longitude “bands” consisting of two 50×50 matrices that repeat the latitude and longitude measurement, respectively. Concatenating the vegetation bands and these two pseudo-bands together gives us an overall auxiliary dimension of $5 \times 50 \times 50$.

UNet. Since our auxiliary information takes the form of 50×50 bands, we need a model architecture that can reconstruct these bands in order to implement the aux-outputs and the In-N-Out models. With this in mind, we utilize a similar UNet architecture that Wang et al. [49] use on the same Cropland dataset. While the UNet was originally proposed by Ronneberger et al. [39] for image segmentation, it can be easily modified to perform image-to-image translation. In particular, we remove the final 1×1 convolutional layer and sigmoid activation that was intended for binary segmentation and replace them with a single convolutional layer whose output dimension matches that of the auxiliary information. In our case, the last convolutional layer has an output dimension of 5 to reconstruct the 3 vegetation bands and (lat,lon) coordinates.

To perform image-level binary classification with the UNet, we also replace the final 1×1 convolutional layer and sigmoid activation, this time with a global average pool and a single linear layer with an output dimension of 1. During training we apply a sigmoid activation to this linear layer’s output to produce a binary class probability, which is then fed into the binary cross entropy loss function.

Aux-inputs model. Since the original RGB input image is $3 \times 50 \times 50$, we can simply concatenate the auxiliary info alongside the original image to produce an input of dimensions $8 \times 50 \times 50$ to feed into the UNet.

Aux-outputs model. The modification of the traditional UNet architecture in order to support auxiliary outputs for Cropland is described in the above UNet section. We additionally add a tanh activation function to squeeze the model’s output values to the range $[-1, 1]$ (the same range as the images). We train the model to learn the auxiliary bands via pixel-wise regression using the mean squared error loss.

In-N-Out model. We found that the finetuning phase of the In-N-Out algorithm experienced wild fluctuations in loss and would not converge when using the hyperparameters listed at the top of this section. To encourage the model to converge and fit the training set, we decreased the Adam learning rate to 0.0001 and doubled the batch size to 512.

Repeated self-training. For the additional round of self-training, we initialize training and pseudolabel all unlabeled data with the In-N-Out model. Following [27], we employ additional regularization when doing self training by adding dropout with probability 0.8.

F.2 Landcover

Our Landcover dataset comes from NASA’s MODIS Surface Reflectance product, which is made up of measurements from around the globe taken by the Terra satellite [48]. In each trial, we use about 16k labeled examples from non-African locations, 203k unlabeled examples (with 150k in-distribution examples), 9266 examples each for validation and in-distribution test, and 4552 OOD test examples. We trained with SGD + momentum (0.9) on all models for 400 epochs with a cosine learning rate schedule. We used learning rate 0.1 for all models that were not pre-trained, and learning rate 0.01 for models that were already pre-trained. Results are reported over 5 trials, and $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ was chosen using the validation set.

1D CNN While Convolutional Neural Networks are most commonly associated with the groundbreaking success of 2D-CNNs on image-based tasks, the 1-dimensional counterparts have also found success in various applications [25]. Because the measurements from the MODIS satellite are not images but instead scalar-valued time series data, we can use a 1D CNN with 7 channels, one for each of the 7 MODIS sensors.

NDVI The normalized difference vegetation index (NDVI) is a remote sensing measurement indicating the presence of live green vegetation. It has been shown to be a useful predictor in landcover-related tasks [11, 10, 31], so we choose to include it in our models as well. NDVI can be computed from the RED and NIR bands of the MODIS sensors via the equation

$$\text{NDVI} = (\text{NIR} - \text{RED}) / (\text{NIR} + \text{RED}). \quad (106)$$

We include NDVI along with the 7 other MODIS bands to give us input dimensions of 46×8 .

ERA5 It is a reasonable hypothesis that having additional climate variables such as soil type or precipitation could be useful for a model in inferring the underlying landcover class. To this end we incorporate features from the ERA5 climate dataset as our auxiliary information [4]. The specific variables we include are soil type, temperature, precipitation rate, precipitation total, solar radiation, and cloud cover. For each MODIS point we find its nearest ERA5 neighbor based on their latitude and longitude in order to pair the datasets together.

The ERA5 measurements are monthly averages, which means the readings are at a different frequency than that of the 8-day MODIS time series. We upsample the ERA5 signal using the `scipy.signal.resample` method, which uses the FFT to convert to the frequency domain, adds extra zeros for upsampling to the desired frequency, and then transforms back into the time domain.

Landcover classes. The Landcover dataset has a total of 16 landcover classes, with a large variance in the individual class counts. To ensure our model sees enough examples of each class, we filtered the dataset to include just 6 of the most populous classes: savannas, woody_savannas, croplands, open_shrublands, evergreen_broadleaf_forests, and grasslands.

Aux-inputs model. We concatenate the resampled ERA5 readings with the MODIS and NDVI measurements to obtain an input dimension of 46×14 .

Aux-outputs model. Rather than predicting the entire ERA5 time series as an auxiliary output, we instead average the 6 climate variables over the time dimension and predict those 6 means as the auxiliary outputs. We use a smaller learning rate of 0.01 for this pre-trained model.

In-N-Out and Repeated self-training. The In-N-Out model initializes its weights from the aux-outputs model and gets pseudolabeled ID unlabeled data from the aux-inputs model. As with aux-outputs, we use a smaller learning rate of 0.01 for this pre-training model.

For the additional round of self-training, we initialize training and pseudolabel all unlabeled data with the In-N-Out model. Following [27], we employ additional regularization when doing self training by adding dropout with probability 0.5. We found that with dropout, we need a higher learning rate (0.1) to effectively fit the training set.

F.3 CelebA

For the results in Table 1, we used 7 auxiliary binary attributes included in the CelebA dataset: ['Bald', 'Bangs', 'Mustache', 'Smiling', '5_o_Clock_Shadow', 'Oval_Face', 'Heavy_Makeup']. These attributes tend to be fairly robust to our distribution shift (not hat vs. hat) — if the person has a 5 o'clock shadow, the person is likely a man. We use a subset of the CelebA dataset with 2000 labeled examples, 30k in-distribution unlabeled examples, 3000 OOD unlabeled examples, and 1000 validation, in-distribution test, and OOD test examples each. The backbone for all models is a ResNet-18 [16] which takes a CelebA image downsized to 64×64 and outputs a binary gender prediction. All models are trained for 25 epochs using SGD with cosine learning rate decay, initial learning rate 0.1, and early stopped with an in-distribution validation set. The gender ratios in the in-distribution and OOD set are balanced to 50-50.

Aux-inputs model. We incorporate the auxiliary inputs by first training a baseline model \hat{f}_{bs} from images to output logit, then training a logistic regression model on the concatenated features $[\hat{f}_{bs}(x); z]$ where z are the auxiliary inputs. We sweep over L2 regularization hyperparameters $C = [0.1, 0.5, 1.0, 5.0, 10.0, 20.0, 50.0]$ and choose the best with respect to an in-distribution validation set.

Aux-outputs model. During pretraining, the model trains on the 7-way binary classification task of predicting the auxiliary information. Then, the model is finetuned on the gender classification task without auxiliary information.

In-N-Out and repeated self-training. For In-N-Out models with repeated self-training, we pseudolabeled all the unlabeled data using the In-N-Out model and did one round of additional self-training. Following [27], we employ additional regularization when doing self training by adding dropout with probability 0.8. We also reduced the learning rate to 0.05 to improve the training dynamics.

Adding auxiliary inputs one-by-one. In Figure 5, we generate a random sequence of 15 auxiliary inputs and add them one-by-one to the model, retraining with every new configuration. We use the following auxiliary information: 'Young', 'Straight_Hair', 'Narrow_Eyes', 'Mouth_Slightly_Open', 'Blond_Hair', '5_o_Clock_Shadow', 'Big_Nose', 'Oval_Face', 'Chubby', 'Attractive', 'Blurry', 'Goatee', 'Heavy_Makeup', 'Wearing_Necklace', and 'Bushy_Eyebrows'.

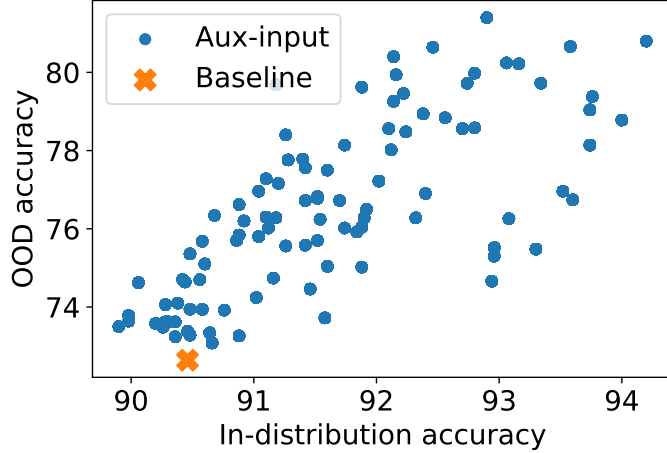


Figure 4: Correlation ($r^2 = 0.52$) between in-distribution accuracy and out-of-distribution (OOD) accuracy when adding 1 to 15 random auxiliary features as input in CelebA.

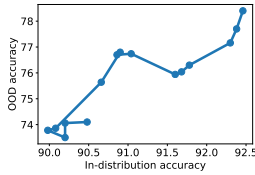


Figure 5: In-distribution vs. OOD accuracy when sequentially adding a random set of 15 auxiliary inputs one-by-one. While more auxiliary inputs generally improves both in-distribution and OOD accuracy, some in-distribution gains can hurt OOD.

	ID Test Acc	OOD Test Acc
Only in-distribution	69.73 \pm 0.51	57.73 \pm 1.58
Only OOD	69.92 \pm 0.41	59.28 \pm 1.01
Both	70.07 \pm 0.46	59.84 \pm 0.98

Table 2: Ablation study on the use of in-distribution vs. OOD unlabeled data in pre-training models on Landcover, where unlabeled sample size is standardized (much smaller than Table 1). Using OOD unlabeled examples are important for gains in OOD accuracy (%). Results are shown with 90% error intervals over 5 trials.

Correlation between in-distribution and OOD accuracy. In Figure 4, we sample 100 random sets of auxiliary inputs of sizes 1 to 15 and train 100 different aux-inputs models using these auxiliary inputs. We plot the in-distribution and OOD accuracy for each model, showing that there is a significant correlation between in-distribution and OOD accuracy in CelebA, supporting results on standard datasets [37, 53, 41]. Each point in the plot is an averaged result over 5 trials.

G Additional Experiments

G.1 Choice of auxiliary inputs matters

We find that the choice of auxiliary inputs affects the tradeoff between ID and OOD performance significantly, and thus is important to consider for problems with distribution shift. While Figure 4 shows that auxiliary inputs tend to simultaneously improve ID and OOD accuracy in CelebA, our theory suggests that in the worst case, there should be auxiliary inputs that worsen OOD accuracy. Indeed, Figure 5 shows that when taking a random set of 15 auxiliary inputs and adding them sequentially as auxiliary inputs, there are instances where an extra auxiliary input improves in-distribution but hurts OOD accuracy. In cropland prediction, we compare using location coordinates and vegetation data as auxiliary inputs with only using vegetation data. The model with locations achieves the best ID performance, improving almost 1% in-distribution over the baseline with only RGB. Without locations, the ID accuracy is similar to the baseline but the OOD accuracy improves by 1.5%. In this problem, location coordinates help with in-distribution interpolation, but the model fails to extrapolate on locations.

G.2 OOD unlabeled data is important for pretraining

We compare the role of in-distribution vs. OOD unlabeled data in pretraining. Table 2 shows the results of using only in-distribution vs. only OOD vs. a balanced mix of unlabeled examples for pretraining on the Landcover dataset, where unlabeled sample size is standardized across the models (by reducing to the size of the smallest set, resulting in 4x less unlabeled data). Using only in-distribution unlabeled examples does not improve OOD accuracy, while having only OOD unlabeled examples does well

both in-distribution and OOD since it also has access to the labeled in-distribution data. For the same experiment in cropland prediction, the differences were not statistically significant, perhaps due to the smaller geographic shift (across states in cropland vs. continents in landcover).