Connect Later: Improving Fine-tuning for Robustness with Targeted Augmentations

Helen Qu [1] **Sang Michael Xie** [2]

Abstract

Models trained on a labeled source domain often generalize poorly when deployed on an out-of-distribution (OOD) target domain. In the domain adaptation setting where unlabeled target data is available, self-supervised pretraining (e.g., contrastive learning or masked autoencoding) is a promising method to mitigate this performance drop. Pretraining depends on generic data augmentations (e.g., cropping or masking) to learn representations that generalize across domains, which may not work for all distribution shifts. In this paper, we show on real-world tasks that standard fine-tuning after pretraining does not consistently improve OOD error over simply training from scratch on labeled source data. To better leverage pretraining for distribution shifts, we propose the Connect Later framework, which fine-tunes the model with *targeted augmentations* designed with knowledge of the shift. Intuitively, pretraining learns good representations within the source and target domains, while fine-tuning with targeted augmentations improves generalization across domains. Connect Later achieves state-of-the-art OOD accuracy while maintaining comparable or better in-distribution accuracy on 4 real-world tasks in wildlife identification (iWildCam-WILDS), tumor detection (Camelyon17-WILDS), and astronomy (AstroClassification, Redshifts).

1. Introduction

In the real world, machine learning models are often deployed on data that differ significantly from training data (Quiñonero-Candela et al., 2009; Koh et al., 2021). We focus on unsupervised domain adaptation (Shimodaira, 2000; Blitzer et al., 2006; Sugiyama et al., 2007), where we have labeled data from a source domain and unlabeled data from a target domain. We aim to learn a model that generalizes well to these out-of-distribution (OOD) target domain inputs. A real-world example is wildlife identification, where the task is to identify animal species from static camera trap images. However, human labels are only available for images from a small subset of these cameras, which may not be representative of the habitats and characteristics of unlabeled camera images.

Pretraining on broad unlabeled data has shown promising results on improving OOD error in real-world problems (Caron et al., 2020; Shen et al., 2022; Radford et al., 2021; Sagawa et al., 2022). In particular, contrastive pretraining has been shown to learn representations that transfer well across domains (Shen et al., 2022; HaoChen et al., 2022). In contrast to conventional domain adaptation methods that focus on learning domain-invariant features (Ganin et al., 2016; Kang et al., 2019; Tzeng et al., 2017; Saenko et al., 2010; Sun et al., 2016; Hoffman et al., 2018), contrastive pretraining learns representations that are not domain-invariant, but instead decompose the class and domain information, facilitating transfer across domains (Shen et al., 2022). A favorable decomposition depends on the generic data augmentations used during contrastive pretraining to align representations across domains. Intuitively, augmented (e.g. masked or cropped) source and target inputs should be more likely to look similar if they are from the same class (e.g., cropping out the face of a lion in different habitats) than from different classes (e.g., no body parts of elephants and lions are alike). However, these generic augmentations may not be suitable for all distribution shifts.

In this paper, we find on real-world benchmarks that standard fine-tuning after contrastive pretraining is not always effective for improving OOD error over purely supervised learning from scratch with labeled source data (Section 3). On the other hand, supervised learning with *targeted augmentations* (Gao et al., 2023) designed for the distribution shift improves OOD error over the supervised learning base-

[*]Equal contribution [1]Department of Physics and Astronomy, University of Pennsylvania [2]Department of Computer Science, Stanford University. Correspondence to: Helen Qu <helenqu@sas.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

line on all datasets without access to any target unlabeled data. Thus, pretraining does not always learn representations that transfer across domains with standard fine-tuning.

To better leverage pretraining for domain adaptation, we propose the Connect Later framework (Figure 1): after pretraining with generic augmentations, fine-tune with targeted augmentations (Section 4). Intuitively, pretraining learns good representations within each domain, while targeted augmentations incorporate domain knowledge to improve generalization across domains. Through both empirical and theoretical examples, we show that Connect Later generalizes well to the target domain even in scenarios where pretraining alone produces minimal OOD performance improvements. We provide a general methodology for constructing these targeted augmentations by matching augmented inputs to the target distribution on a feature space where the domains differ.

We evaluate our framework on 4 real-world datasets: wildlife identification (iWildCam-WILDS, Beery et al., 2020; Sagawa et al., 2022), tumor detection (Camelyon17-WILDS, Bandi et al., 2018; Sagawa et al., 2022) and 2 astronomical time series tasks, AstroClassification and Redshifts, which we curate from The PLAsTiCC team et al. (2018). In Section 5, we show that Connect Later improves OOD performance over standard fine-tuning or supervised learning with targeted augmentations across all datasets. Although our understanding stems from contrastive learning, we empirically apply Connect Later to better leverage pretrained representations from both masked autoencoding and contrastive learning. Connect Later achieves the state-of-the-art on three benchmarks, improving OOD accuracy on AstroClassification by 3% (Boone, 2019), iWildCam-WILDS with ResNet-50 by 0.9%, and Camelyon17-WILDS with DenseNet121 by 1.1%. We also contribute the Redshifts dataset, on which Connect Later outperforms the best baseline by 11% relative improvement.

2. Setup

We consider a prediction problem from an input space $\mathcal{X}$ to a label space $\mathcal{Y}$, where $\mathcal{Y} = \{1, \dots, k\}$ for classification and $\mathcal{Y} \in \mathbb{R}$ for regression.

Domain adaptation. Let P_S and P_T be the source and target input distributions over $\mathcal{X}$, respectively. We consider unsupervised domain adaptation, where we have access to source inputs $x \sim P_S$, with corresponding labels $y \in \mathcal{Y}$ sampled from the label distribution $p^*(\cdot \mid x)$, along with unlabeled target inputs sampled from the target distribution P_T. Let the unlabeled distribution $P_U = \beta P_S + (1-\beta) P_T$ be a mixture of the source and target, where $\beta \in [0, 1]$. In practice, P_U may also be a broader unlabeled distribution.

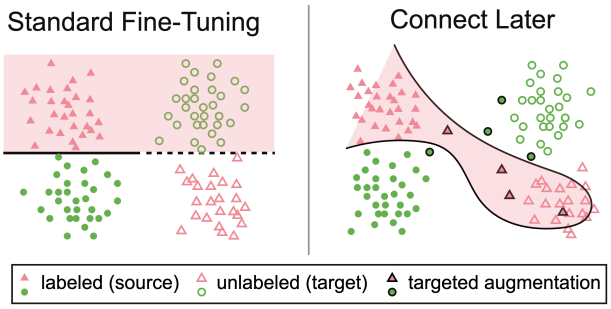

Figure 1. Overview of the Connect Later framework applied to a toy binary classification problem with two domains (filled and unfilled points), showing the representations from contrastive pretraining in $\mathbb{R}^2$. **(Left)** After contrastive pretraining with generic augmentations, the classes within each domain are linearly separable in representation space. Since the domains are far apart in input space, generic augmentations may misalign the pretrained representations. In this case, a classifier (with a linearly extrapolating decision boundary, dashed and solid line) learned on labeled source data will misclassify the target data. **(Right)** Connect Later employs targeted augmentations (filled points with black border) designed with knowledge of the distribution shift to connect the domains better, resulting in a classifier that generalizes well to the target domain.

The goal is to learn a model $f : \mathcal{X} \to \mathcal{Y}$ that minimizes error on the target domain $L_T(f) = \mathbb{E}_{x \sim P_T, y \sim p^*(\cdot|x)}[\ell(f(x), y)]$. For example, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the 0-1 loss in classification and squared loss in regression.

Augmentations. Augmented inputs $x' \in \mathcal{X}$ are drawn from an augmentation distribution $\mathcal{A}(\cdot|x)$, given an input $x \in \mathcal{X}$. Training with augmented inputs is often used to improve robustness (Hendrycks et al., 2019; 2020) and is crucial to contrastive pretraining (Caron et al., 2020; Shen et al., 2022; Devlin et al., 2019). In this work, we define two distinct augmentation distributions, $\mathcal{A}_{\text{pre}}$ and $\mathcal{A}_{\text{ft}}$, for the pretraining and fine-tuning steps, respectively. Typically, the pretraining augmentations $\mathcal{A}_{\text{pre}}$ are generic transformations, such as random cropping in vision or masking in NLP (Caron et al., 2020; Chen et al., 2020; He et al., 2020; Radford et al., 2021; Shen et al., 2022; He et al., 2022; Devlin et al., 2019). Fine-tuning augmentations $\mathcal{A}_{\text{ft}}$ have not been studied extensively and are typically also generic or simply the identity transformation (Sagawa et al., 2022; Devlin et al., 2019).

Contrastive pretraining for domain adaptation. Contrastive pretraining for domain adaptation consists of two steps: self-supervised pretraining on unlabeled data, then supervised fine-tuning on labeled source data (Shen et al., 2022). For simplicity below, we consider the population objectives. Contrastive learning aims to learn an encoder which maps augmented views of the same input to similar

features ("positive pairs") and views of different inputs to dissimilar features ("negative pairs"), according to some distance metric. Formally, let $S_+(x, x^+) = \mathbb{E}_{\bar{x}\sim P_U}[\mathcal{A}_{\text{pre}}(x \mid \bar{x})\mathcal{A}_{\text{pre}}(x^+ \mid \bar{x})]$ be the distribution over positive pairs, which are augmentations of a single input $\bar{x}$. We pretrain an encoder $\phi : \mathcal{X} \to \mathbb{R}^k$ to minimize the distance d_+ between positive pair embeddings and maximize the distance d_- between negative pair embeddings:

$$\mathcal{L}_{\text{pretrain}}(\phi) = \mathbb{E}_{(x,x^+)\sim S_+}[d_+(\phi(x), \phi(x^+))] - \mathbb{E}_{x,x'\sim P_U}[d_-(\phi(x), \phi(x'))]. \tag{1}$$

The output of the pretraining step is a pretrained encoder $\widehat{\phi} = \arg\min_\phi \mathcal{L}_{\text{pretrain}}(\phi)$.

Fine-tuning then learns a prediction head $h : \mathbb{R}^k \to \mathbb{R}^n$ (for regression, we let $n = 1$) on top of the pretrained encoder using labeled source data with the objective

$$\mathcal{L}_{\text{ft}}(h) = \mathbb{E}_{x\sim P_S, y\sim p^*(\cdot|x), x'\sim\mathcal{A}_{\text{ft}}(\cdot|x)}[\text{loss}_{\text{ft}}(h(\widehat{\phi}(x')),\ y;\ \theta)] \tag{2}$$

where $\text{loss}_{\text{ft}} : \mathbb{R}^n \times \mathcal{Y} \to \mathbb{R}$ is a fine-tuning objective such as softmax cross entropy loss for classification or squared error for regression. The learned head is $\widehat{h} = \arg\min_h \mathcal{L}_{\text{ft}}(h)$. In practice, we jointly fine-tune the head h and the encoder $\widehat{\phi}$.

Standard fine-tuning. We refer to **standard fine-tuning** as the pretraining+fine-tuning procedure where $\mathcal{A}_{\text{ft}}(x' \mid x) = 1$ if $x' = x$ (no fine-tuning augmentations). In our experiments, the standard fine-tuning procedure is linear probing then fine-tuning (LP-FT) (Kumar et al., 2022), which has been shown to improve ID and OOD performance over vanilla fine-tuning. In LP-FT, we first learn a linear predictor on top of frozen pretrained features before fine-tuning all the parameters jointly.

ERM with augmentations. As a baseline, we consider empirical risk minimization (ERM) with data augmentation, which optimizes the fine-tuning objective (Equation 2) on labeled source data with randomly initialized parameters. In this paper, we refer to **ERM** as the instantiation where $\mathcal{A}_{\text{ft}}(x' \mid x) = 1$ if $x' = x$ (no augmentations) and **ERM + targeted augmentations** as the instantiation with $\mathcal{A}_{\text{ft}}$ that is designed with knowledge of the distribution shift.

3. Pretraining may not improve OOD performance

We compare ERM and standard fine-tuning on two benchmark datasets, iWildCam-WILDS (wildlife species identification) and Camelyon17-WILDS (tumor detection). In Table 1, we show that standard fine-tuning on a model pretrained using SwAV contrastive learning (Caron et al., 2020) makes minimal gains over ERM on iWildCam-WILDS (46.4 → 46.4 ID, 30.4 → 31.2 OOD) compared to Camelyon17-WILDS (89.3 → 92.3 ID, 65.2 → 91.4 OOD). This result runs contrary to prior work demonstrating that contrastive pretraining is an effective domain adaptation method (Caron et al., 2020; Shen et al., 2022; Radford et al., 2021; Sagawa et al., 2022). We hypothesize that the generic pretraining augmentations connect the domains better for some tasks and distribution shifts than others.

Table 1. Contrastive pretraining with standard fine-tuning substantially improves OOD performance for Camelyon17-WILDS but is not very effective for iWildCam-WILDS. Results are averaged over 15 trials for iWildCam-WILDS and 20 trials for Camelyon17-WILDS, and we report the 95% confidence intervals on each mean estimate.

	iWildCam (Macro F1, ↑)		Camelyon17 (Avg Acc, ↑)	
	ID Test	OOD Test	ID Val	OOD Test
ERM	46.4 ± 0.5	30.4 ± 0.6	89.3 ± 0.9	65.2 ± 1.1
Standard fine-tuning	46.4 ± 0.8	31.2 ± 0.6	92.3 ± 0.2	91.4 ± 0.9

Simple example with misaligned connectivity structure. To understand this phenomenon, we provide a simple binary classification example of when contrastive pretraining fails for domain adaptation, following a similar augmentation graph construction to Shen et al. (2022), in Appendix E. When the connectivity structure misaligns the source and target domains, such that examples from the same class are less "connected" than examples from different classes across the domains, a linear classifier trained on these pretrained representations will not transfer from source to target. This could happen, for example, when the source and target are far apart in input space and connectivity is low between examples from the same class across different domains.

3.1. Robustness gains from pretraining depend on dataset connectivity

To better understand why contrastive pretraining performs differently on these two datasets, we empirically evaluate the connectivity measures for iWildCam-WILDS and Camelyon17-WILDS. We follow Shen et al. (2022) and work in the augmentation graph setting, where nodes are inputs and edge weights are the positive-pair probabilities given by S_+. We define connectivity between a class-domain pair $((y_1, d_1), (y_2, d_2))$ under four scenarios:

$$\begin{cases} \rho & y_1 = y_2, d_1 = d_2 \quad \text{(same class, same domain)} \\ \alpha & y_1 = y_2, d_1 \neq d_2 \quad \text{(same class, different domain)} \\ \beta & y_1 \neq y_2, d_1 = d_2 \quad \text{(different class, same domain)} \\ \gamma & y_1 \neq y_2, d_1 \neq d_2 \quad \text{(different class and domain)} \end{cases}, \tag{3}$$

Table 2. Empirically estimated connectivity measures for iWildCam-WILDS and Camelyon17-WILDS. From Shen et al. (2022), contrastive pretraining theoretically learns transferable representations for UDA when both across-domain (α) and across-class (β) connectivity is greater than across-both (γ). In iWildCam-WILDS, $\gamma > \beta$, violating the condition, while Camelyon17-WILDS satisfies the condition.

	α	β	γ
iWildCam-WILDS	0.116	0.071	0.076
Camelyon17-WILDS	0.16	0.198	0.152

where each value is an average edge weight over the edges that satisfy each case. Shen et al. (2022) show in simple augmentation graphs that contrastive pretraining theoretically learns transferable representations when $\alpha > \gamma$ and $\beta > \gamma$, and that the ratios $\frac{\alpha}{\gamma}$ and $\frac{\beta}{\gamma}$ empirically correlate well with OOD accuracy. Intuitively, the pretraining augmentations are less likely to change both the domain and class of an input than changing just domain or just class.

Empirical evaluations of connectivity. We empirically evaluate the connectivity measures for iWildCam-WILDS and Camelyon17-WILDS following Shen et al. (2022). Using augmented inputs from 2 class-domain pairs, we train a binary classifier to predict the class-domain pair of each input, and interpret the test error of the classifier as an estimate for connectivity. We average each connectivity value over 15 class-domain pairs (see Appendix D for details). Our results, summarized in Table 2, show that iWildCam-WILDS connectivity measures violate the condition for contrastive pretraining in the UDA setting, since across-both connectivity $>$ across-class ($\gamma > \beta$). This finding is consistent with our observation that contrastive pretraining is far less effective for iWildCam-WILDS compared to Camelyon17-WILDS, and further underscores the need for domain adaptation methods that correct the misaligned connectivity structure.

4. Connect Later: Pretrain First, Targeted Augmentations Later

Even when generic augmentations applied during pretraining misalign the connectivity structure, the pretrained representations are still useful since the classes are linearly separable *within* each domain. How do we leverage these pretrained representations when they may not transfer well across domains? In this work, we propose the Connect Later framework (Figure 1):

1. Pretrain on unlabeled data with generic augmentations as in Equation 1, producing a pretrained encoder $\widehat{\phi}$.
2. Design a targeted augmentation $\mathcal{A}_{\text{ft}}$ (discussed below) and use augmented source data to fine-tune the pretrained encoder $\widehat{\phi}$ jointly with a prediction head h as in Equation 2.

While our intuition about pretraining for domain adaptation stems from Shen et al. (2022), we show that applying targeted augmentations at fine-tuning time is sufficient for good generalization to the target domain even when the pretrained representations transfer across domains poorly. This allows us to reuse pretrained models for multiple downstream tasks.

Simple example where Connect Later achieves 0 OOD error. In our simple binary classification example in Appendix E, we show that when the connectivity structure is misaligned, both standard fine-tuning with contrastive pretraining and ERM + targeted augmentations have high OOD error, while Connect Later achieves 0 OOD error. In this setting, ERM with targeted augmentations is unable to achieve 0 OOD error since some target inputs are "unreachable" via targeted augmentations of source inputs. The pretraining step in Connect Later uses unlabeled target data to learn representations where label information from source data can propagate to all target inputs.

4.1. Real-world examples of targeted augmentations

We design targeted augmentations for 4 real-world tasks: wildlife identification, tumor detection, and astronomical time-series classification and redshift prediction. We show examples from the source, augmented, and target datasets for these tasks in Figure 2.

Wildlife species classification (iWildCam-WILDS). For iWildCam-WILDS (Beery et al., 2020; Sagawa et al., 2022), the task is to identify the wildlife species from static camera trap images. These cameras are placed in a wide variety of environments, which all have unique habitat conditions (e.g., African savannah vs. tropical rainforest) and camera characteristics (e.g., angles, resolutions).

- **Source:** 243 camera traps
- **Target:** 48 unseen camera traps
- **Targeted Augmentation:** We augment the labeled dataset with the Copy-Paste Same Y algorithm, which uses image segmentation to copy-paste the animal onto different background images from cameras that have observed the same species (Gao et al., 2023).
- **Task:** 182-class wildlife species classification

Tumor detection (Camelyon17-WILDS). The task in Camelyon17-WILDS (Bandi et al., 2018) is to classify

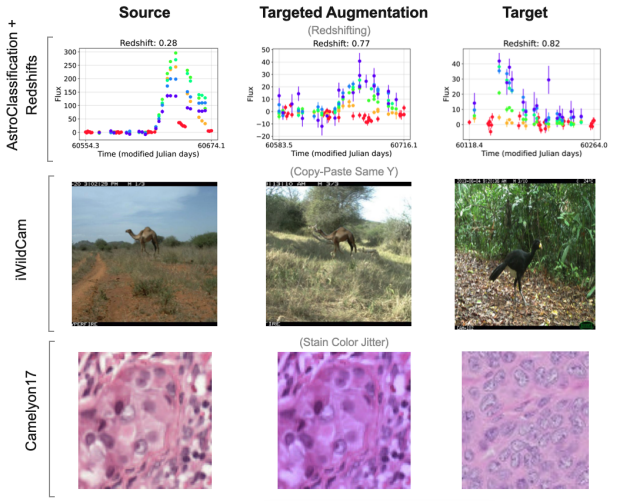

Figure 2. Examples from the source dataset (left), an augmented version of the source example (middle), and the target dataset (right) for each of our tasks. (**Top row**) The ASTROCLASSIFICATION and REDSHIFTS tasks focus on time-varying astronomical objects observed in multiple wavelength ranges, plotted here as a multicolored time-series with each color corresponding to the wavelength range of the measurement. The redshifting augmentation simulates placing source objects at a higher redshift to better match the redshift distribution of the target dataset. The flux errors and flux values of the augmented example (middle) show much better resemblance to the target example. (**Middle row**) We randomize the habitat background by applying the Copy-Paste Same Y augmentation for IWILDCAM-WILDS (IWILDCAM-WILDS image examples shown here are from Gao et al. (2023)). (**Bottom row**) Stain Color Jitter alters the overall color of source images in CAMELYON17-WILDS to improve performance on images from unseen hospitals.

whether a patch of a histopathology slide contains a tumor. These slides are contributed from multiple hospitals, which use different stain colors and also vary in distributions of patient cancer stage.

- **Source:** Hospitals 1-3.
- **Target:** Hospitals 4 and 5.
- **Targeted Augmentation:** We augment the labeled dataset with the Stain Color Jitter algorithm, which jitters the color of the slide image in the hematoxylin and eosin staining color space (Tellez et al., 2018).
- **Task:** Determine if a slide contains a tumor.

Astronomical object classification (ASTROCLASSIFICATION). Astronomical object classification (Boone, 2019; Allam Jr. & McEwen, 2022) involves predicting the object type (e.g., type II supernova) from a time series of an object's brightness at multiple wavelengths. Expert labeling is only available for nearby objects, which are brighter and have different properties than distant objects (see Appendix A.1 for details).

- **Source:** Time-series of bright, nearby objects with expert labels
- **Target:** Time-series of all observed objects from the telescope, often faint and distant (higher redshift). Follow-up observation, which is required for expert labeling, is too expensive for these objects.
- **Targeted Augmentation:** We augment the labeled dataset by redshifting each object, i.e., simulating its observed properties as if it were further away (details in Appendix B.2).
- **Task:** 14-class astronomical object classification

Redshift regression (REDSHIFTS). Similar to object type, redshift information is also available only for bright, nearby objects. We predict the scalar redshift value of each object and minimize mean squared error. REDSHIFTS is a new dataset that we contribute as part of this work.

- **Source:** Time-series of bright, nearby labeled objects.
- **Target:** Time-series of all observed objects from the telescope, often faint and distant (higher redshift).
- **Targeted Augmentation:** Redshifting (same as ASTROCLASSIFICATION, Appendix B.2).
- **Task:** Redshift regression

4.2. Designing targeted augmentations

Targeted augmentations offer the opportunity to incorporate domain knowledge to improve generalization performance. We provide a general methodology for designing targeted augmentations based on matching the target distribution on a feature space:

1. Identify a feature space $\mathcal{Z}$. We assume that we can label $z \in \mathcal{Z}$ for each input and that the source and target domains largely differ on this feature space. One such example is the space of spurious, domain-dependent features (e.g., camera angle or resolution for IWILDCAM-WILDS), which is the approach followed by Gao et al. (2023).
2. Fit a transformed feature distribution $\hat{p_T}(z'|z)$ to the target feature distribution.
3. Create a transformation distribution $T(x'|x, z')$ where x' is the augmented version of x with $z = z'$. In this paper, we define T with domain knowledge.

4. Given an input x, generate augmentations by sampling a new feature z' from $\hat{p_T}(z' \mid z)$, then sampling an augmentation from $T(x'|x, z')$. The resulting targeted augmentation probabilities are $\mathcal{A}_{\text{ft}}(x' \mid x) = \sum_{z'} T(x' \mid x, z')\hat{p_T}(z' \mid z)$.

Targeted augmentation example. We follow the procedure outlined above to design a targeted augmentation for AstroClassification and Redshifts (see Appendix B.2 for further details).

1. The source and target domains have different redshift distributions, so we identify this scalar feature as z.

2. We roughly fit the target redshift distribution within a reasonable range of the original redshift z, such that $\hat{p_T}(z' \mid z)$ is distributed as loguniform$(0.95z,\ \min(1.5(1+z)-1,\ 5z))$, following Boone (2019).

3. We define a transformation distribution $T(x'|x, z')$, where x is a time-series of flux values at multiple wavelengths and z' is a new redshift value to transform to. We first fit a Gaussian process that models x as a function of time and wavelength. Given z', we rescale the timestamps and wavelengths of the original input to account for the physical effects of the new redshift value. Then, we sample $\tilde{x'}$ from the Gaussian process at these new timestamps and wavelengths. Finally, we produce the transformed input x' by scaling the flux values to account for z'.

4. We sample z' from $\hat{p_T}(z' \mid z)$ and then sample augmentations x' from $T(x'|x, z')$.

5. Experiments

We empirically test Connect Later with contrastive pretraining (iWildCam-WILDS, Camelyon17-WILDS) as well as pretraining with masked autoencoding (AstroClassification, Redshifts) to demonstrate Connect Later as a general fine-tuning method. We note that masked autoencoding has been linked to contrastive learning in Zhang et al. (2022), which shows that the masked autoencoding objective upper bounds the contrastive loss between positive pairs – thus, masked autoencoding implicitly aligns the positive pairs induced by the masking augmentations.

Training procedure. For iWildCam-WILDS, we use a ResNet-50 model pretrained on unlabeled ImageNet data with SwAV contrastive learning (Caron et al., 2020). We use a DenseNet121 pretrained on unlabeled data from Sagawa et al. (2022) with SwAV for Camelyon17-WILDS. We pretrain with masked autoencoding for AstroClassification and Redshifts by masking 60% of observations from each light curve (Appendix C). The same pretrained model is used for both tasks to demonstrate the reusability of pretrained features. We fine-tune the pretrained models with linear probing then fine-tuning (LP-FT, Kumar et al., 2022), which has been shown to improve OOD performance.

Baselines. We evaluate our framework against three baselines: ERM, ERM+targeted augs, and standard fine-tuning. We include Avocado (Boone, 2019), the state-of-the-art model for AstroClassification. We also include a self-training baseline for AstroClassification and Redshifts, which has been shown to perform well on some real-world datasets (Sagawa et al., 2022). For the self-training baseline, we pseudo-label the target dataset with a trained ERM+targeted augs model, then combine with the labeled source dataset and apply the targeted augmentation for training. We include additional domain adaptation baselines for iWildCam-WILDS and Camelyon17-WILDS: domain-adversarial neural networks (DANN, Ganin et al., 2016), correlation alignment (CORAL, Sun et al., 2016), Noisy Student (Xie et al., 2020b), and ICON[1].

5.1. Main results

Tables 3 and 4 compare the results of Connect Later with baseline methods. Connect Later outperforms all baselines on the OOD metric while maintaining comparable or better ID performance and achieves state-of-the-art performance on iWildCam-WILDS by 0.8% OOD for ResNet-50, Camelyon17-WILDS by 1.1% OOD for DenseNet121, and AstroClassification by 3%.

Connect Later improves OOD performance when standard fine-tuning is minimally effective for UDA. On iWildCam-WILDS, standard fine-tuning minimally improves in OOD performance, while ERM+targeted augmentations improves by 6% ID and OOD over both ERM and standard fine-tuning. Connect Later improves over both standard fine-tuning (by 6.8%) and ERM+targeted augs (by 0.9%) in OOD performance, indicating that the pretrained representations as well as the targeted augmentations are both important for OOD performance. Connect Later achieves a new state-of-the-art performance for ResNet-50 on the iWildCam-WILDS benchmark.

When standard fine-tuning is effective, Connect Later still produces additional performance gains. On Camelyon17-WILDS, AstroClassification, and Redshifts, Connect Later still outperforms all variants even though standard fine-tuning already produces significant gains over ERM. For Camelyon17-WILDS, standard fine-tuning improves substantially over ERM in OOD average accuracy (26.2%). ERM+targeted augs outperforms

[1] https://github.com/a-tea-guy/ICON

Table 3. ID and OOD accuracy (%) for AstroClassification and RMSE for Redshifts of each method. Results are averaged over 5 trials and rows with means within 1 STD of the best mean are bolded.

	AstroClassification		Redshift	
	ID Test Acc (↑)	OOD Acc (↑)	ID Test RMSE (↓)	OOD RMSE (↓)
ERM	71.59 ± 1.10	61.26 ± 1.10	0.274 ± 0.016	0.320 ± 0.009
Standard fine-tuning	78.84 ± 0.97	67.84 ± 0.70	$\mathbf{0.246 \pm 0.015}$	0.277 ± 0.004
ERM + targeted augs	68.75 ± 0.95	67.54 ± 0.32	0.310 ± 0.006	0.286 ± 0.007
Self-Training	77.72 ± 0.59	65.15 ± 0.67	0.304 ± 0.010	0.289 ± 0.003
Avocado (Boone, 2019)	-	77.40	-	-
Connect Later	$\mathbf{80.54 \pm 1.20}$	$\mathbf{79.90 \pm 0.60}$	$\mathbf{0.256 \pm 0.005}$	$\mathbf{0.247 \pm 0.005}$

Table 4. ID and OOD performance for each method on iWildCam-WILDS and Camelyon17-WILDS. Results are averaged over 15 trials for iWildCam-WILDS and 20 trials for Camelyon17-WILDS, and we report 95% confidence intervals on each mean estimate. Rows with means within 1 interval of the best mean are bolded.

	iWildCam (Macro F1, ↑)		Camelyon17 (Avg Acc, ↑)	
	ID Test	OOD Test	ID Val	OOD Test
ERM	46.4 ± 0.5	30.4 ± 0.6	89.3 ± 0.9	65.2 ± 1.1
Standard fine-tuning	46.4 ± 0.8	31.2 ± 0.6	92.3 ± 0.2	91.4 ± 0.9
ERM + targeted augs	$\mathbf{51.4 \pm 0.6}$	36.1 ± 0.7	96.7 ± 0.0	90.5 ± 0.4
DANN (Sagawa et al., 2022)	48.5 ± 3.2	31.9 ± 1.6	86.1 ± 1.3	64.5 ± 1.2
CORAL (Sagawa et al., 2022)	40.5 ± 1.6	27.9 ± 0.5	92.3 ± 0.7	62.3 ± 1.9
Noisy Student (Sagawa et al., 2022)	47.5 ± 1.0	32.1 ± 0.8	-	-
ICON	50.6 ± 1.3	34.5 ± 1.4	90.1 ± 0.4	93.8 ± 0.3
Connect Later	$\mathbf{51.7 \pm 0.8}$	$\mathbf{36.9 \pm 0.7}$	$\mathbf{98.5 \pm 0.0}$	$\mathbf{94.9 \pm 0.4}$

standard fine-tuning in ID accuracy by 4.4%, but does not improve OOD. Connect Later sets a new state-of-the-art on Camelyon17-WILDS with DenseNet121, improving on the best ID performance by 1.8% (ERM+targeted augs) and OOD performance by 1.1% (ICON).

For AstroClassification, standard fine-tuning also performs significantly better than ERM: 7% ID, 6.5% OOD. ERM+targeted augs underperforms in ID accuracy compared to ERM (-2.8%) and standard fine-tuning (-9.9%), likely due to the strong targeted augmentations. However, OOD accuracy of ERM+targeted augs is competitive with standard fine-tuning, outperforming ERM. Connect Later outperforms the best baseline, standard fine-tuning, by 12% OOD and 2% ID. The ID accuracy outperforms both standard fine-tuning and ERM+targeted augs, showing a complementary benefit between pretraining and targeted augmentations. Connect Later sets a new state-of-the-art OOD performance on AstroClassification by 3% over Avocado, a heavily tuned random forest model with expert-designed features (Boone, 2019).

Redshifts results are similar to AstroClassification, with standard fine-tuning significantly improving over ERM in both ID (7% relative) and OOD (13% relative) RMSE. Connect Later outperforms the best baseline variant, standard fine-tuning, by 0.03 RMSE (11% relative) with comparable ID error.

Connect Later improves OOD performance for CLIP fine-tuning. We additionally evaluated the effectiveness of Connect Later for CLIP ViT-L/14 (Radford et al., 2021) on the iWildCam-WILDS dataset (Table 5). Standard fine-tuning improves substantially over both ERM and ERM + targeted augs, likely due to CLIP's internet-scale pretraining dataset as well as the practical importance of pretraining for ViT performance. However, Connect Later still delivers additional gains over standard fine-tuning (0.5% ID, 1.4% OOD).

Table 5. ID and OOD Macro F1 results for fine-tuning CLIP ViT-L on iWildCam-WILDS. Results are averaged over 15 trials and we report 95% confidence intervals on each mean estimate. Rows with means within 1 interval of the best mean are bolded.

	ID Test	OOD Test
ERM	22.6 ± 0.6	7.8 ± 0.2
Standard fine-tuning	$\mathbf{55.3 \pm 1.4}$	42.8 ± 0.9
ERM + targeted augs	23.8 ± 0.7	8.4 ± 0.4
Connect Later	$\mathbf{55.8 \pm 0.8}$	$\mathbf{44.2 \pm 0.9}$

Other baselines. DANN, CORAL, and Noisy Student did not produce competitive OOD average accuracy for either iWildCam-WILDS or Camelyon17-WILDS. ICON is the best baseline for Camelyon17-WILDS OOD average accuracy and is outperformed only by Connect Later. For AstroClassification and Redshifts, self-training improves both ID and OOD performance compared to ERM but underperforms standard fine-tuning in both domains.

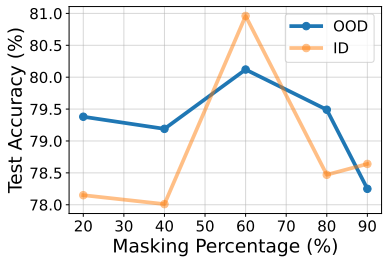

Figure 3. On the AstroClassification task, Connect Later is relatively robust to pretraining masking percentage both ID and OOD, but 60% masking performs best out of the percentages we tested.

5.2. Ablations

We performed ablations on the model size, strength of pretraining augmentations (masking percentage for masked autoencoding), and LP-FT on AstroClassification. We find that downstream performance is quite robust to masking percentage, while scaling up model size and LP-FT improve performance for pretrained models.

Model scale. We tested Connect Later with a larger model ($\sim 3\times$ the parameters of our model, 21M $\rightarrow$ 69M), and find that the larger model produces higher ID and OOD accuracy (Table 6). This suggests that scaling up the model is a promising way to further improve performance with Connect Later.

Strength of pretraining augmentations (masking percentage). We vary the strength of pretraining augmentations with the MAE objective, as augmentation strength is parameterized solely by masking percentage. We tested pretraining masking percentages $\{20, 40, 60, 80, 90\}\%$ with the same masking strategy (replace 10% of masked indices with random values from the time-series, another 10% are kept unchanged, and 80% are replaced with the mask token, 0). We show the ID and OOD test accuracy of each variant in Figure 3. Both ID and OOD performance peak at 60% masking, although the performance of Connect Later is quite robust to the masking percentage. All masking percentages improve on OOD performance over standard fine-tuning or ERM with targeted augmentations. Even the strongest augmentations (90% masking) did not improve OOD performance over weaker augmentations. We hypothesize that strong generic augmentations may indiscriminately increase the connectivity between all source and target examples, including examples from different classes that should not be strongly connected.

Linear probing then fine-tuning. Kumar et al. (2022) showed that linear probing (with fixed neural embeddings) and then fine-tuning (LP-FT) the entire model improves both ID and OOD performance. Intuitively, full fine-tuning with a randomly initialized linear probe can destroy the pretrained features, and training the linear probe first mitigates this. We test LP-FT against FT only (all model weights are fine-tuned) with the Connect Later model and the ERM+targeted augs baseline. We find that LP-FT improves OOD accuracy by 0.9% over FT only when applied to Connect Later on AstroClassification (Table 7). On the other hand, LP-FT decreased OOD accuracy by 1.4% when applied to ERM+targeted augs, which uses random initialization (no pretraining). As a result, we use LP-FT on pretrained models but not on ERM or ERM+targeted augs.

Table 6. Scaling up model size of Connect Later produces improvements in both ID and OOD performance on the AstroClassification task.

Number of Parameters	ID Acc (↑)	OOD Acc (↑)
21M (default)	78.47	79.49
69M	80.38	80.55

Table 7. Linear probing (LP) in addition to fine-tuning (FT) hurts performance for the ERM+targeted augs model but improves performance for Connect Later (tested on the AstroClassification task).

	Connect Later		ERM+targeted augs	
	ID Acc(↑)	OOD Acc(↑)	ID Acc(↑)	OOD Acc(↑)
FT only	78.07	78.6	77.88	68.43
LP-FT	78.47	79.49	65.68	67.07

6. Discussion and Related Work

Augmentations for pretraining. Data augmentations (e.g., cropping or masking) are vital to semi- and self-supervised learning objectives. Reconstructing a masked or noised input has been shown to produce useful pretrained representations across multiple modalities (Devlin et al., 2019; Lewis et al., 2020; He et al., 2022; Raffel et al., 2019; Chen et al., 2020; He et al., 2020; Caron et al., 2020). In contrastive learning, models are trained to distinguish augmented "views" of the same input from views of a different input (Chen et al., 2020; Caron et al., 2020; He et al., 2020). Our results demonstrate that though pretraining with generic augmentations alone produces inconsistent OOD performance across datasets, fine-tuning with targeted augmentations is able to better leverage these pretrained representations.

Augmentations for robustness. Data augmentation has been used to improve model robustness to label-independent changes (e.g. translation or rotation in vision) (Hendrycks et al., 2019; Rebuffi et al., 2021; Ng et al., 2020). Existing augmentation strategies rely on generic perturbations that

aim to increase the diversity of inputs (e.g., Simard et al., 2003; Krizhevsky et al., 2012; Cubuk et al., 2019; 2020; DeVries & Taylor, 2017; Zhang et al., 2017), though prior work has shown that the type of data augmentations matters for performance (Chen et al., 2020; Xie et al., 2020a). Augmentations have also been leveraged in the self-training paradigm, which improves generalization to unseen data by training on the pseudo-labeled full dataset (Xie et al., 2020b; Sohn et al., 2020; Yang et al., 2021). We show that a self-training baseline with pseudo-labels from an ERM+targeted augs model does not outperform Connect Later, indicating that pretraining is important for robustness gains. Connect Later exposes targeted augmentations as a design interface for improving robustness with knowledge of the distribution shift, while still leveraging pretrained representations.

Targeted augmentations. In domain shift problems, Gao et al. (2023) show that targeted augmentations designed with knowledge of the distribution shift outperform generic, or even target-aware (e.g. CutMix, Yun et al. (2019)), augmentations on unseen data. Gao et al. (2023) consider the domain generalization setting, in which the target dataset is unknown. We consider targeted augmentations in the domain adaptation setting, where we can model the target distribution with the unlabeled target data. In this setting, targeted augmentations provide the opportunity to naturally incorporate domain knowledge about the dataset and distribution shift. In this work, we provide a general methodology for the design of such augmentations and show that targeted augmentations better leverage pretrained representations for complementary gains in OOD performance. Certain aspects of the design process, such as the selection of feature space z and transformation distribution T could be learned from the unlabeled data itself, which we leave for future work.

7. Conclusion

We show that pretraining with generic augmentations is not a panacea for all distribution shifts and tasks, and does not deliver consistent gains over supervised learning on labeled source data. Pure supervised learning, however, does not use the unlabeled data or produce reusable representations. Connect Later allows for better leverage of pretrained representations for OOD performance by applying targeted augmentations at fine-tuning time.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Allam Jr., T. and McEwen, J. D. Paying attention to astronomical transients: Introducing the time-series transformer for photometric classification, 2022.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Beery, S., Cole, E., and Gjoka, A. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

Blitzer, J., McDonald, R., and Pereira, F. Domain adaptation with structural correspondence learning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

Boone, K. Avocado: Photometric classification of astronomical transients with gaussian process augmentation. *The Astronomical Journal*, 158(6):257, dec 2019. doi: 10.3847/1538-3881/ab5182. URL https://doi.org/10.3847%2F1538-3881%2Fab5182.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9912–9924, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 702–703, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics (ACL)*, pp. 4171–4186, 2019.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Vishnu, M., and Vardhan, H. ASTROMER. *Astronomy & Astrophysics*, 670:A54, feb 2023. doi: 10.1051/0004-6361/202243928. URL https://doi.org/10.1051%2F0004-6361%2F202243928.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17, 2016.

Gao, I., Sagawa, S., Koh, P. W., Hashimoto, T., and Liang, P. Out-of-domain robustness via targeted augmentations. In *International Conference on Machine Learning (ICML)*, 2023.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *arXiv preprint arXiv:2106.04156*, 2021.

HaoChen, J. Z., Wei, C., Kumar, A., and Ma, T. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *arXiv*, 2022.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., Angel, J. R. P., Angeli, G. Z., Ansari, R., Antilogus, P., Araujo, C., Armstrong, R., Arndt, K. T., Astier, P., Aubourg, É., Auza, N., Axelrod, T. S., Bard, D. J., Barr, J. D., Barrau, A., Bartlett, J. G., Bauer, A. E., Bauman, B. J., Baumont, S., Bechtol, E., Bechtol, K., Becker, A. C., Becla, J., Beldica, C., Bellavia, S., Bianco, F. B., Biswas, R., Blanc, G., Blazek, J., Blandford, R. D., Bloom, J. S., Bogart, J., Bond, T. W., Booth, M. T., Borgland, A. W., Borne, K., Bosch, J. F., Boutigny, D., Brackett, C. A., Bradshaw, A., Brandt, W. N., Brown, M. E., Bullock, J. S., Burchat, P., Burke, D. L., Cagnoli, G., Calabrese, D., Callahan, S., Callen, A. L., Carlin, J. L., Carlson, E. L., Chandrasekharan, S., Charles-Emerson, G., Chesley, S., Cheu, E. C., Chiang, H.-F., Chiang, J., Chirino, C., Chow, D., Ciardi, D. R., Claver, C. F., Cohen-Tanugi, J., Cockrum, J. J., Coles, R., Connolly, A. J., Cook, K. H., Cooray, A., Covey, K. R., Cribbs, C., Cui, W., Cutri, R., Daly, P. N., Daniel, S. F., Daruich, F., Daubard, G., Daues, G., Dawson, W., Delgado, F., Dellapenna, A., de Peyster, R., de Val-Borro, M., Digel, S. W., Doherty, P., Dubois, R., Dubois-Felsmann, G. P., Durech, J., Economou, F., Eifler, T., Eracleous, M., Emmons, B. L., Fausti Neto, A., Ferguson, H., Figueroa, E., Fisher-Levine, M., Focke, W., Foss, M. D., Frank, J., Freemon, M. D., Gangler, E., Gawiser, E., Geary, J. C., Gee, P., Geha, M., Gessner, C. J. B., Gibson, R. R., Gilmore, D. K., Glanzman, T., Glick, W., Goldina, T., Goldstein, D. A., Goodenow, I., Graham, M. L., Gressler, W. J., Gris, P., Guy, L. P., Guyonnet, A., Haller, G., Harris, R., Hascall, P. A., Haupt, J., Hernandez, F., Herrmann, S., Hileman, E., Hoblitt, J., Hodgson, J. A., Hogan, C., Howard, J. D., Huang, D., Huffer, M. E., Ingraham, P., Innes, W. R., Jacoby, S. H., Jain, B., Jammes, F., Jee, M. J., Jenness, T., Jernigan, G., Jevremović, D., Johns, K., Johnson, A. S., Johnson, M. W. G., Jones, R. L., Juramy-Gilles, C., Jurić, M., Kalirai, J. S., Kallivayalil, N. J., Kalmbach, B., Kantor, J. P., Karst, P., Kasliwal, M. M., Kelly, H., Kessler, R., Kinnison, V., Kirkby, D., Knox, L., Kotov, I. V., Krabbendam, V. L., Krughoff, K. S., Kubánek, P., Kuczewski, J., Kulkarni, S., Ku, J., Kurita, N. R., Lage, C. S., Lambert, R., Lange, T., Langton, J. B., Le Guillou, L., Levine, D., Liang, M., Lim, K.-T., Lintott, C. J., Long, K. E., Lopez, M., Lotz, P. J., Lupton, R. H., Lust, N. B., MacArthur, L. A., Mahabal, A., Mandelbaum, R., Markiewicz, T. W., Marsh, D. S., Marshall, P. J., Marshall, S., May, M., McKercher, R., McQueen, M., Meyers, J., Migliore, M., Miller, M., Mills, D. J., Miraval, C., Moeyens, J., Moolekamp, F. E., Monet, D. G., Moniez, M., Monkewitz, S., Montgomery, C., Morrison, C. B., Mueller, F., Muller, G. P., Muñoz Arancibia, F., Neill, D. R., Newbry, S. P., Nief, J.-Y., Nomerotski, A., Nordby, M., O'Connor, P., Oliver, J., Olivier, S. S., Olsen, K., O'Mullane, W., Ortiz, S., Osier, S., Owen, R. E., Pain, R., Palecek, P. E., Parejko, J. K., Parsons, J. B., Pease, N. M., Peterson, J. M., Peterson, J. R., Petravick, D. L., Libby Petrick, M. E., Petry, C. E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P. A., Plante, R., Plate, S., Plutchak, J. P., Price, P. A., Prouza, M., Radeka,

V., Rajagopal, J., Rasmussen, A. P., Regnault, N., Reil, K. A., Reiss, D. J., Reuter, M. A., Ridgway, S. T., Riot, V. J., Ritz, S., Robinson, S., Roby, W., Roodman, A., Rosing, W., Roucelle, C., Rumore, M. R., Russo, S., Saha, A., Sassolas, B., Schalk, T. L., Schellart, P., Schindler, R. H., Schmidt, S., Schneider, D. P., Schneider, M. D., Schoening, W., Schumacher, G., Schwamb, M. E., Sebag, J., Selvy, B., Sembroski, G. H., Seppala, L. G., Serio, A., Serrano, E., Shaw, R. A., Shipsey, I., Sick, J., Silvestri, N., Slater, C. T., Smith, J. A., Smith, R. C., Sobhani, S., Soldahl, C., Storrie-Lombardi, L., Stover, E., Strauss, M. A., Street, R. A., Stubbs, C. W., Sullivan, I. S., Sweeney, D., Swinbank, J. D., Szalay, A., Takacs, P., Tether, S. A., Thaler, J. J., Thayer, J. G., Thomas, S., Thornton, A. J., Thukral, V., Tice, J., Trilling, D. E., Turri, M., Van Berg, R., Vanden Berk, D., Vetter, K., Virieux, F., Vucina, T., Wahl, W., Walkowicz, L., Walsh, B., Walter, C. W., Wang, D. L., Wang, S.-Y., Warner, M., Wiecha, O., Willman, B., Winters, S. E., Wittman, D., Wolff, S. C., Wood-Vasey, W. M., Wu, X., Xin, B., Yoachim, P., and Zhan, H. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873(2): 111, March 2019. doi: 10.3847/1538-4357/ab042c.

Kang, G., Jiang, L., Yang, Y., and Hauptmann, A. G. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4893–4902, 2019.

Kessler, R., Bernstein, J. P., Cinabro, D., Dilday, B., Frieman, J. A., Jha, S., Kuhlmann, S., Miknaitis, G., Sako, M., Taylor, M., and Vanderplas, J. SNANA: A Public Software Package for Supernova Analysis. *Proceedings of the Astronomical Society of the Pacific*, 121(883):1028, September 2009. doi: 10.1086/605984.

Kessler, R., Narayan, G., Avelino, A., Bachelet, E., Biswas, R., Brown, P. J., Chernoff, D. F., Connolly, A. J., Dai, M., Daniel, S., Stefano, R. D., Drout, M. R., Galbany, L., Gonzá lez-Gaitán, S., Graham, M. L., Hložek, R., Ishida, E. E. O., Guillochon, J., Jha, S. W., Jones, D. O., Mandel, K. S., Muthukrishna, D., O'Grady, A., Peters, C. M., Pierel, J. R., Ponder, K. A., Prša, A., Rodney, S., and and, V. A. V. Models and simulations for the photometric LSST astronomical time series classification challenge (PLAsTiCC). *Publications of the Astronomical Society of the Pacific*, 131(1003):094501, jul 2019. doi: 10.1088/1538-3873/ab26f1. URL https://doi.org/10.1088%2F1538-3873%2Fab26f1.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1097–1105, 2012.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Association for Computational Linguistics (ACL)*, 2020.

Li, Y., Si, S., Li, G., Hsieh, C.-J., and Bengio, S. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems*, 34:15816–15829, 2021.

Nakoneczny, S., Bilicki, M., Pollo, A., Asgari, M., Dvornik, A., Erben, T., Giblin, B., Heymans, C., Hildebrandt, H., Kannawadi, A., et al. Photometric selection and redshifts for quasars in the kilo-degree survey data release 4. *Astronomy & Astrophysics*, 649:A81, 2021.

Ng, N., Cho, K., and Ghassemi, M. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1268–1283, 2020.

Qu, H. and Sako, M. Photo-zSNthesis: Converting Type Ia Supernova Lightcurves to Redshift Estimates via Deep Learning. *Astrophysical Journal*, 954(2):201, September 2023. doi: 10.3847/1538-4357/aceafa.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. A. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226, 2010.

Sagawa, S., Koh, P. W., Lee, T., Gao, I., Xie, S. M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., Beery, S., David, E., Stavness, I., Guo, W., Leskovec, J., Saenko, K., Hashimoto, T. B., Levine, S., Finn, C., and Liang, P. Extending the WILDS benchmark for unsupervised adaptation. In *International Conference on Learning Representations (ICLR)*, 2022.

Shen, K., Jones, R., Kumar, A., Xie, S. M., HaoChen, J. Z., Ma, T., and Liang, P. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, 2022.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.

Simard, P. Y., Steinkraus, D., and Platt, J. C. Best practices for convolutional neural networks applied to visual document analysis. *International Conference on Document Analysis and Recognition*, 2:958–964, 2003.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv*, 2020.

Sugiyama, M., Krauledat, M., and Muller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research (JMLR)*, 8:985–1005, 2007.

Sun, B., Feng, J., and Saenko, K. Return of frustratingly easy domain adaptation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.

Tellez, D., Balkenhol, M., Otte-H "oller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9): 2126–2136, 2018.

The PLAsTiCC team, au2, T. A. J., Bahmanyar, A., Biswas, R., Dai, M., Galbany, L., Hložek, R., Ishida, E. E. O., Jha, S. W., Jones, D. O., Kessler, R., Lochner, M., Mahabal, A. A., Malz, A. I., Mandel, K. S., Martínez-Galarza, J. R., McEwen, J. D., Muthukrishna, D., Narayan, G., Peiris, H., Peters, C. M., Ponder, K., Setzer, C. N., Collaboration, T. L. D. E. S., Transients, T. L., and Collaboration, V. S. S. The photometric lsst astronomical time-series classification challenge (plasticc): Data set, 2018.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. *arXiv*, 2020b.

Yang, Q., Wei, X., Wang, B., Hua, X.-S., and Zhang, L. Interactive self-training with mean teachers for semi-supervised object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5937–5946, 2021. doi: 10.1109/CVPR46437.2021.00588.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zhang, H., Cissé, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *ICLR*, 2017.

Zhang, Q., Wang, Y., and Wang, Y. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35: 27127–27139, 2022.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

A. Additional Dataset Details

A.1. AstroClassification, Redshifts Datasets

The ASTROCLASSIFICATION and REDSHIFTS datasets were adapted from the 2019 Photometric LSST Astronomical Time-Series Classification Challenge (The PLAsTiCC team et al., 2018) [2]. This diverse dataset contains 14 types of astronomical time-varying objects, simulated using the expected instrument characteristics and survey strategy of the upcoming Legacy Survey of Space and Time (LSST Ivezić et al., 2019) conducted at the Vera C. Rubin Observatory. It includes two overall categories of time-series objects: *transients*, short-lived events such as supernovae, and *variable* sources, those with fluctuating brightness such as pulsating stars. Specifically, the dataset includes the following transients: type Ia supernovae (SNIa), SNIax, SNIa-91bg, SNIbc, SNII, superluminous supernovae (SLSN), tidal disruption events (TDE), and single lens microlensing events (μLens-Single); and the following variable objects: active galactic nuclei (AGN), Mira variables, eclipsing binary systems (EB), and RR Lyrae (RRL).

Millions of potential new objects are discovered per observing night, and important metadata such as object type, redshift, or other physical parameters, require astronomers to take time-intensive *spectra* of each object. Spectra are a granular brightness vs. wavelength measurement at a single point in time, and are typically only taken for bright, nearby objects which require less exposure time than faint, faraway objects. The vast majority of discovered objects, however, will not have spectra but instead a time series of imaging data taken in 6 broad wavelength ranges, or *photometric bands*. The time-varying behavior of these objects in these coarse wavelength bands does offer important clues about these physical parameters, but expert interpretation of spectra are traditionally required for confident labeling. Thus, our labeled training data for both ASTROCLASSIFICATION and REDSHIFTS come from the unrepresentative subset of objects with spectra.

In these tasks, we are specifically interested in predicting the object type (e.g. type II supernova) and the cosmological redshift of objects in the unlabeled dataset. *Cosmological redshift* is a proxy for distance in the universe, and an important piece of metadata for understanding an object's physical processes as well as other applications, such as estimating the expansion rate of the universe with type Ia supernovae. The redshift prediction task has been studied for individual object types, such as quasars (Nakoneczny et al., 2021) and type Ia supernovae (Qu & Sako, 2023), but we consider a more realistic set of multiple object types.

Problem Setting. The task is to predict object type for ASTROCLASSIFICATION (redshift for REDSHIFTS) from time-series of object brightness. The input x consists of flux measurements and associated uncertainties at times $\boldsymbol{t}$ and photometric band that each measurement was taken in $\boldsymbol{b}$: $\{F(t_i, b_j)\}_{i=1,j=1}^{T,W}, \{F_{\text{err}}(t_i, b_j)\}_{i=1,j=1}^{T,W}$. For this work, we map each $b \in \boldsymbol{b}$ to the central wavelength of the b band, which we denote $\boldsymbol{w}$. The domain d is binary, corresponding to whether the object has a spectrum (and thus a label). The labels y are available only for objects with spectra, and are one of 14 types of astronomical time-varying objects for ASTROCLASSIFICATION (redshift of the object for REDSHIFTS). We seek to optimize performance on the unlabeled data, which are generally fainter and further away than the labeled subset. We evaluate on these examples as well as held-out examples from the labeled subset.

Data. The training set of 7,846 objects is designed to emulate a sample of objects with spectra and thus biased toward brighter, more nearby objects compared to the test set of 3,492,888 objects. A random subset of 10,000 test set objects was selected for evaluation.

1. **Source:** 6,274 objects
2. **ID Test**: 782 objects
3. **OOD Test:** 10,000 objects

All data were simulated with the SuperNova ANAlysis (SNANA, Kessler et al., 2009) software library. Further details about the astrophysical models and LSST instrument characteristics used in the simulation can be found in Kessler et al. (2019).

[2] https://zenodo.org/record/2539456

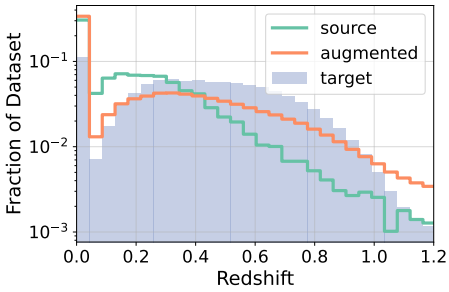

Figure 4. Redshift distributions of source, augmented, and target datasets for the AstroClassification and Redshifts tasks.

B. Data Augmentations

B.1. Generic Augmentations for Pretraining

AstroClassification and Redshifts. For the AstroClassification and Redshifts datasets, we randomly mask a subset of the input sequence using the masked language modeling paradigm introduced by (Devlin et al., 2019). Given an unlabeled input sequence x, a training input x' can be generated by randomly masking elements of x while the associated label y consists of the original, unmasked values. The model is trained to use contextual information (unmasked elements) to successfully reconstruct most of the sequence. From our ablation experiments, we find that a masking percentage of 60% produces the best downstream results. We follow an existing implementation for astronomical time-series (Donoso-Oliva et al., 2023) and set 80% of the masked elements to 0, replace 10% with a random element from the sequence, and keep the remaining 10% unchanged.

iWildCam and Camelyon17. For iWildCam-WILDS, we use a ResNet-50 model pretrained on ImageNet with SwAV, a contrastive learning algorithm (Caron et al., 2020). For Camelyon17-WILDS, we use a DenseNet121 pretrained with SwAV on the unlabeled Camelyon17-WILDS dataset from Sagawa et al. (2022). SwAV uses random cropping augmentations of different resolutions.

B.2. Targeted Augmentations for Fine-Tuning

Redshifting for AstroClassification and Redshifts. The OOD test set of the AstroClassification and Redshifts datasets have many more high redshift objects than the source dataset, leading us to adopt an augmentation scheme to alleviate this shift. Figure 4 shows the redshift distributions of the source, augmented, and target datasets. Redshifting places each object at a new redshift and recomputes its light curve sampling, fluxes, and flux uncertainties accordingly. This augmentation algorithm was adapted from Boone (2019).

An input $\boldsymbol{X} \in \mathbb{R}^{T\times W}$ is a multivariate time series of flux values at specified times and observed wavelengths, $\{F(t_i, w_j)\}_{i=1,j=1}^{T,W}$. We also have $\boldsymbol{X}_{\text{err}} \in \mathbb{R}^{T\times W}$, representing the flux errors corresponding to each element of $\boldsymbol{X}$. We denote the elements of $\boldsymbol{X}'_{\text{err}}$ by $\{F_{\text{err}}(t_i, w_j)\}_{i=1,j=1}^{T,W}$.Our goal is to model $F, F_{\text{err}} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ at a new chosen redshift, z', to produce augmented inputs $\boldsymbol{X}', \boldsymbol{X}'_{\text{err}}$.

- We first construct a distribution from which to sample the new redshift, taking into account the current redshift of the object z_{orig} as well as the target redshift distribution. We then sample a new redshift, $z' \sim \text{loguniform}(0.95z_{\text{orig}},\ \min(1.5(1+z_{\text{orig}})-1,\ 5z_{\text{orig}}))$.
- We fit a Gaussian process (GP) model for F with training observations $\boldsymbol{X}$ queried at the training input values $(\boldsymbol{t}, \boldsymbol{w})$, and denote the predictive mean and variance of the GP as F', F'_{err}.
- Given the new redshift value z', we rescale the timestamps and wavelengths of the original observations to account for the physical effects of the new redshift value: $\boldsymbol{t}_{\text{new}} = \frac{1+z'}{1+z_{\text{orig}}}\boldsymbol{t}$, $\boldsymbol{w}_{\text{new}} = \frac{1+z'}{1+z_{\text{orig}}}\boldsymbol{w}$. We also randomly drop out 10% as well as a large swath of $(\boldsymbol{t}_{\text{new}}, \boldsymbol{w}_{\text{new}})$ to simulate distinct observing seasons (telescope observing only occurs in the winter).

- We obtain GP predictions at test inputs $\{F'(t_{\text{new},i}, w_{\text{new},j})\}_{i=1,j=1}^{T,W}$, $\{F'_{\text{err}}(t_{\text{new},i}, w_{\text{new},i})\}_{i=1,j=1}^{T,W}$ and scale them by the log ratio of the new and original distances:

$$\tilde{\boldsymbol{X}}' = 10^{0.4(d(z')-d(z_{\text{orig}}))}\{F'(t_{\text{new},i}, w_{\text{new},j})\}_{i=1,j=1}^{T,W},$$

$$\tilde{\boldsymbol{X}}'_{\text{err}} = 10^{0.4(d(z')-d(z_{\text{orig}}))}\{F'_{\text{err}}(t_{\text{new},i}, w_{\text{new},j})\}_{i=1,j=1}^{T,W},$$

  where $d(z)$ is the distance corresponding to redshift z.

- We roughly model the observational noise of the telescope from the target data as a function of wavelength and sample $\epsilon \in \mathbb{R}^W$ from it. We define

$$\boldsymbol{X}' = \{\tilde{\boldsymbol{X}}'_{:,j} + \epsilon_j\}_{j=1}^{W}, \boldsymbol{X}'_{\text{err}} = \left\{\sqrt{\tilde{\boldsymbol{X}}'^2_{\text{err},:,j} + \epsilon_j^2}\right\}_{j=1}^{W}.$$

- We model the observational capabilities of the telescope to ensure that our augmented input $\boldsymbol{X}'$, $\boldsymbol{X}'_{\text{err}}$ does not fall below the threshold of detection. We "accept" an augmented input $\boldsymbol{X}'$, $\boldsymbol{X}'_{\text{err}}$ if the signal-to-noise ratio (SNR) of at least two observations is over 5, i.e. $\text{SNR}(\boldsymbol{X}'_{i,j}, \boldsymbol{X}'_{\text{err},i,j}) \geq 5$ for at least 2 of $i \in \{1, ..., T\}, j \in \{1, ..., W\}$. We define $\text{SNR}(x, x_{\text{err}}) = \frac{|x|}{x_{\text{err}}}$.

Copy-Paste (Same Y) for iWildCam. This augmentation strategy randomizes the backgrounds of wildlife images to reduce the model's dependence on these spurious features for species classification. Specifically, a segmentation mask is applied to each image to separate the animal from the background, and the animal is "copy-pasted" into a new background from a camera that has observed that animal species. This was the best performing augmentation strategy from Gao et al. (2023).

Stain Color Jitter for Camelyon17. This augmentation, originally from Tellez et al. (2018), alters the pixel values of the slide images to emulate different staining procedures used by different hospitals. The augmentation uses a pre-specified Optical Density (OD) matrix to project images from RGB space to a three-channel hematoxylin, eosin, and DAB space before applying a random linear combination. This was the best performing augmentation strategy from Gao et al. (2023).

C. Experimental Details

AstroClassification and Redshifts. For AstroClassification and Redshifts, we pretrain with a masked autoencoding objective:

$$\mathcal{L}_{\text{MAE}}(\phi) = \mathbb{E}_{x\sim P_U, x'\sim\mathcal{A}_{\text{pre}}(\cdot|x)}[(\phi(x') - x)^2] \tag{4}$$

We use an encoder-only Informer model (Zhou et al., 2021) with 8 encoder layers of 12 attention heads each. The model hidden dimension was chosen to be 768 and the layer MLPs have hidden dimension 256. Due to the 2-dimensional position data (each element of the time-series has an associated time and photometric band/wavelength) and irregular sampling of our dataset, we train a positional encoding based on learnable Fourier features following Li et al. (2021). We also select a random window of length 300 from each example (and zero-pad examples with fewer than 300 observations) to produce inputs of uniform shape. We perform pretraining with a batch size of 256 and learning rate 1e-4 (selected from 1e-3 $\sim$ 1e-6) for 75,000 steps. We finetune the pretrained model with linear probing for 20,000 steps (for pretrained models only) and learning rate 1e-4, then fine-tuning for 10,000 steps at learning rate of 4e-5. We increase the learning rate for models without pretraining to 1e-4 for FT. The Redshifts task uses LP learning rate of 5e-4 and FT learning rate of 1e-4. We decrease the learning rate per step with a linear scheduler.

iWildCam. For pretraining, we use ResNet-50 pretrained on ImageNet with SwAV (Caron et al., 2020). During fine-tuning, we train all models for 15 epochs with early stopping on OOD validation performance, following Gao et al. (2023). For pretrained models, we also do 10 epochs of linear probing before fine-tuning (LP-FT, Kumar et al., 2022) for 15 epochs, where the linear probe is trained with Adam and the linear probe weights used to initialize the fine-tuning stage is chosen

Table 8. Empirically estimated connectivity measures for iWildCam-WILDS, AstroClassification, and Camelyon17-WILDS. iWildCam-WILDS and AstroClassification results are averaged over 15 randomly selected class-domain pairs, while Camelyon17-WILDS results are averaged over all possible class-domain pairs.

	across-domain	across-class	across-both
iWildCam-WILDS	0.116	0.071	0.076
AstroClassification	0.287	0.159	0.097
Camelyon17-WILDS	0.16	0.198	0.152

with OOD validation performance. To reduce the noise in OOD results, for all methods we select the epoch in the last 5 epochs with the best OOD validation performance and report OOD test results with that version of the model. Following Gao et al. (2023), we allow for 10 hyperparameter tuning runs, where we sample the following hyperparameters independently from the following distributions: the linear probe learning rate ($10^{\text{Uniform}[-3,-2]}$), fine-tuning learning rate ($10^{\text{Uniform}[-5,-2]}$), and probability of applying the augmentation (Uniform$[0.5, 0.9]$) and pick the hyperparameter configuration with the best OOD validation performance. For ERM and ERM+targeted augmentations, we use the tuned hyperparameters from Gao et al. (2023). To decrease the confidence interval due to an outlier seed, the reported performance of Connect Later is averaged over 15 seeds. All other results are averaged over 5 seeds.

Camelyon17. For pretraining, we use DenseNet121 pretrained on the unlabeled Camelyon17-WILDS dataset presented in Sagawa et al. (2022) with SwAV (Caron et al., 2020). During fine-tuning, we train all models for 15 epochs with early stopping on OOD validation performance, following Gao et al. (2023). For pretrained models, we also do 10 epochs of linear probing before fine-tuning (LP-FT, Kumar et al., 2022) for 15 epochs, where the linear probe is trained with Adam and the linear probe weights used to initialize the fine-tuning stage is chosen with OOD validation performance. To reduce the noise in OOD results, for all methods we select the epoch with the best OOD validation performance and report OOD test results with that version of the model. Following Gao et al. (2023), we allow for 10 hyperparameter tuning runs, where we sample the following hyperparameters independently from the following distributions: the linear probe learning rate ($10^{\text{Uniform}[-3,-2]}$), fine-tuning learning rate ($10^{\text{Uniform}[-5,-2]}$), probability of applying the augmentation (Uniform$[0.5, 0.9]$), and augmentation strength (Uniform$[0.05, 0.1]$), and pick the hyperparameter configuration with the best OOD validation performance. All results are averaged over 20 seeds.

D. Empirical Estimates of Connectivity

We empirically estimate connectivity measures for all of the datasets we tested on following the procedure outlined in Appendix D of Shen et al. (2022). Specifically, we train binary classifiers from scratch to predict the class-domain pair of a given input example. We randomly select 15 class-domain pairs for iWildCam-WILDS and AstroClassification, while for Camelyon17-WILDS we use all class-domain pairs since Camelyon17-WILDS is a binary classification task. We label these class-domain examples following Appendix D of Shen et al. (2022) and create a dataset with 80/10/10 train/validation/test split. We train using the same hyperparameters described in Appendix C for 3,000 steps with early stopping on the validation accuracy. Our results are presented in Table 8.

E. Simple construction where Connect Later improves over pretraining or targeted augmentations alone

We give a simple construction for constrastive pretraining based on the construction in Proposition 3 (Appendix A.2) of Shen et al. (2022), where Connect Later improves over pretraining (standard fine-tuning) or targeted augmentations alone.

Data distribution. We consider binary classification with 2 domains. Let $\mathcal{S} = \{x \in \mathcal{X} : d_x = 1\}$ and $\mathcal{T} = \{x \in \mathcal{T} : d_x = 2\}$, and assume that P_S and P_T are uniform over $\mathcal{S}$ and $\mathcal{T}$. The unlabeled distribution for pretraining is the uniform distribution over $\mathcal{X}$. The source domain $\mathcal{S} = \{1, 2\}$ contains 2 points and the target domain $\mathcal{T} = \{3, 4, 5, 6, 7, 8\}$ contains 6 points. For simplicity, we let the labels y_x be a deterministic function of the input x. The label space is $\mathcal{Y} = \{-1, 1\}$. The label for $x \in \{1, 3, 5, 7\}$ is $y_x = 1$ and the label for $x \in \{2, 4, 6, 8\}$ is $y_x = -1$. Only the source data is labeled.

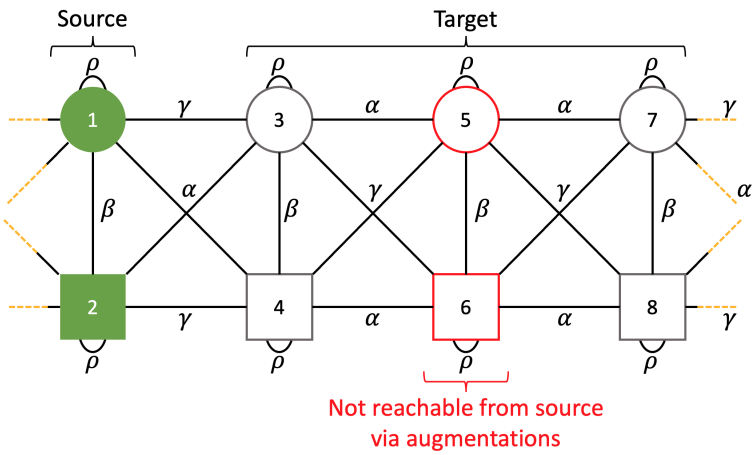

Figure 5. Example distribution of data and augmentations for contrastive learning where Connect Later improves OOD performance over contrastive pretraining+standard fine-tuning and ERM+targeted augmentations. The augmentation graph is similar to Shen et al. (2022) except the edge weights connecting 1,2 and 3,4 are swapped. The shapes represent classes, while the labeled data is shaded in green. The generic augmentation probabilities are marked as edge weights, where we assume that $\alpha > \gamma + \beta$. Here, targeted augmentations which first swap inputs 1 and 2 before applying a generic augmentation help to align the source and target. However, some target inputs are not reachable via augmentations from source inputs. Standard fine-tuning can generalize throughout the target domain, but only in conjunction with targeted augmentations that align the source and target. The orange dotted lines on the far ends connect to each other (the graph wraps around).

ERM with targeted augmentations. ERM with targeted augmentations learns a model on source labeled data. To specialize to this section, the ERM objective is

$$\mathcal{L}_{\text{ERM}}(f) = \mathbb{E}_{x\sim P_S, x'\sim\mathcal{A}_{\text{ft}}(\cdot|x)}[\ell(f(x'), y_x)]. \tag{5}$$

ERM returns a classifier $\widehat{f}_{\text{erm}} \in \arg\min_f \mathcal{L}_{\text{ERM}}(f)$.

Spectral contrastive learning. Following HaoChen et al. (2021) and Shen et al. (2022), we analyze contrastive learning from an augmentation graph perspective, where inputs x are connected via augmentations with edge weights $S_+(x, x')$, which represent the probability of x, x' being a positive pair (augmentations of the same input x). For theoretical analysis, we analyze the spectral contrastive learning objective:

$$\mathcal{L}_{\text{pretrain}}(\phi) = -2\cdot\mathbb{E}_{(x,x^+)\sim S_+}\left[\phi(x)^\top\phi(x^+)\right] + \mathbb{E}_{x,x'\sim P_U}\left[\left(\phi(x)^\top\phi(x')\right)^2\right]. \tag{6}$$

The result of pretraining to optimize the above objective is an encoder $\widehat{\phi} : \mathcal{X} \to \mathbb{R}^k$.

Linear probing (fine-tuning step). Instead of analyzing fine-tuning, we follow Shen et al. (2022) and analyze linear probing on top of the pretrained representations from the encoder. We train a linear model with parameters $B \in \mathbb{R}^{r\times k}$, where r is the number of classes. We minimize the objective:

$$\mathcal{L}(B) = \mathbb{E}_{x\sim P_S}\left[\ell(B\widehat{\phi}(x), y_x)\right] + \eta\|B\|_F^2, \tag{7}$$

where ℓ is the squared loss and we take $y_x \in \mathbb{R}^k$ to be a one-hot encoding of the class label. The resulting classifier is $\widehat{f}(x) = \arg\max_{i\in[r]}(\widehat{B}\widehat{\phi}(x))_i$.

Pretraining augmentations (Figure 5) We define the pretraining augmentation distribution $\mathcal{A}_{\text{pre}}(\cdot \mid x)$ to be

$$\mathcal{A}_{\text{pre}}(x' \mid x) = \begin{cases} \rho' & x = x' \\ \alpha' & \{x', x\} \in \{\{1,4\},\{3,5\},\{5,7\},\{2,5\},\{4,6\},\{6,8\},\{1,8\},\{2,7\} \\ \beta' & \{x', x\} \in \{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\} \\ \gamma' & \{x', x\} \in \{\{1,3\},\{2,4\},\{3,6\},\{4,5\},\{5,8\},\{6,7\},\{1,7\},\{2,8\}\} \end{cases}. \tag{8}$$

Notice that the weight between 1,3 is γ' and the weight between 1,4 is α', and the weights are similarly swapped for 2,4, and 2,5. We assume that ρ', α', β', and γ' are in $(0, 1)$ and are distinct. We also assume that the augmentation probabilities satisfy $\rho' > \max\{\alpha', \beta'\}$ and $\min\{\alpha', \beta'\} > \gamma'$. Following Shen et al. (2022), we can convert these to positive pair probabilities $\rho, \alpha, \beta, \gamma$ with similar properties by renormalizing.

Given the above setting, the following is a simplified form of Proposition 3 from Shen et al. (2022), if we instead use the following augmentation distribution, which swaps the edge weight magnitudes that involve nodes 1 and 2:

$$\mathcal{A}_{\text{prop}}(x' \mid x) = \begin{cases} \rho' & x = x' \\ \alpha' & \{x', x\} \in \{\{1,3\}, \{3,5\}, \{5,7\}, \{2,4\}, \{4,6\}, \{6,8\}, \{1,7\}, \{2,8\} \\ \beta' & \{x', x\} \in \{\{1,2\}, \{3,4\}, \{5,6\}, \{7,8\}\} \\ \gamma' & \{x', x\} \in \{\{1,4\}, \{2,3\}, \{3,6\}, \{4,5\}, \{5,8\}, \{6,7\}, \{1,8\}, \{2,7\}\} \end{cases}. \tag{9}$$

Proposition E.1 (Shen et al. (2022))**.** *With the above construction for the input space $\mathcal{X}$, unlabeled distribution P_U, and data augmentation $\mathcal{A}_{prop}$, for some feature dimension $k \in \mathbb{Z}^+$ a linear probe trained on contrastive pre-trained features achieves 0 target error: $\mathcal{L}_{0-1}(\widehat{f}) = 0$. However, for all $k \in \mathbb{Z}^+$, there exists a minimizer $\widehat{f}_{erm}$ of the ERM objective (with data augmentations according to $\mathcal{A}_{prop}$) that has non-zero error: $\mathcal{L}_{0-1}(\widehat{f}_{erm}) = 1/3$.*

ERM with targeted augmentations can get high OOD error. In general, we proceed by defining the following targeted augmentation, which allows us to reduce to the setting of Proposition E.1:

$$\mathcal{A}_{\text{ft}}(x' \mid x) = \begin{cases} 1 & \{x', x\} \in \{1,4\}, \{2,3\} \\ 1 & x = x' \text{ and } x \notin \{1,2\} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

which transforms input 1 to 4 and the input 2 to 3, while keeping all other inputs the same. Since the ERM with augmentations objective will not contain a term involving inputs 5,6,7, or 8 and thus the prediction on these inputs do not affect the objective, there exists a minimizer of the ERM objective (Equation 5) that predicts the wrong label for inputs 5,6,7,8 and has target error 2/3. This is because these nodes are unreachable via augmentations of the source inputs, and thus the ERM objective can be minimized with any arbitrary prediction on these inputs.

Standard fine-tuning has high OOD error. By Proposition E.1, standard fine-tuning after contrastive pretraining has zero target (OOD) error when the pretraining augmentations do not have swapped edges. By symmetry, standard fine-tuning (contrastive pretraining + linear probing) on our augmentation graph with pretraining augmentations $\mathcal{A}_{\text{pre}}$ outputs the opposite label for all target inputs, resulting in an OOD error of 1. This is because the source and target domains are misaligned in our augmentation graph.

Connect Later achieves zero OOD error. Connect Later applies targeted augmentations $\mathcal{A}_{\text{ft}}$ during the linear probing step (on top of contrastive pretrained representations). This choice of targeted augmentations reduces to the setting of Proposition E.1 where the labeled source domain consists of the inputs 3,4 instead. By the symmetry of the graph and applying Proposition E.1, Connect Later achieves 0 OOD error.